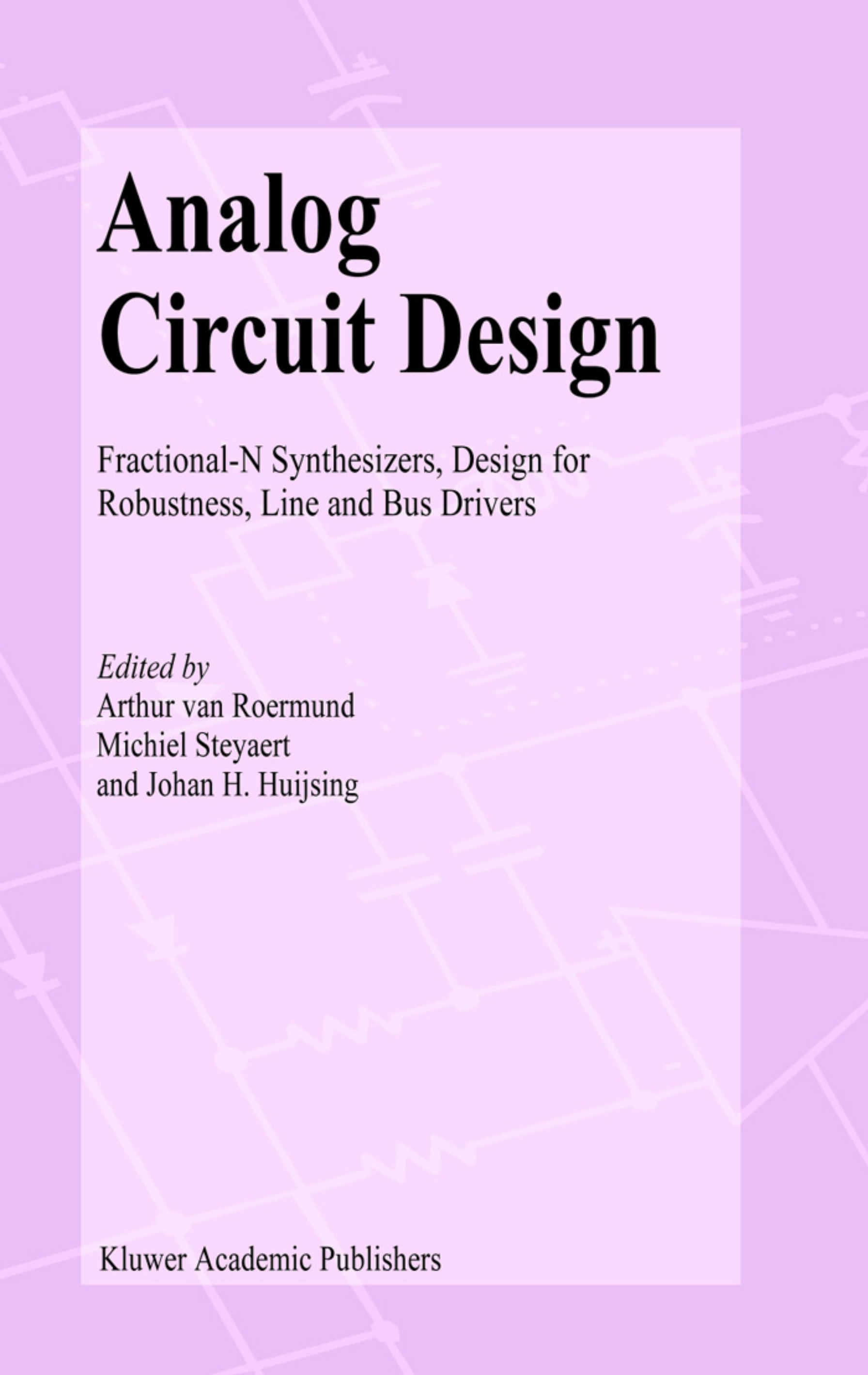


Analog Circuit Design

Fractional-N Synthesizers, Design for Robustness, Line and Bus Drivers

Edited by
Arthur van Roermund
Michiel Steyaert
and Johan H. Huijsing

Kluwer Academic Publishers

VISIT…

LANZAROTE
Caliente.COM

ANALOG CIRCUIT DESIGN

This page intentionally left blank

Analog Circuit Design

Fractional-N Synthesizers, Design for Robustness, Line and Bus Drivers

Edited by

Arthur van Roermund
Eindhoven University of Technology, The Netherlands

Michiel Steyaert
KU Leuven, Belgium

and

Johan H. Huijsing
Delft University of Technology, The Netherlands

KLUWER ACADEMIC PUBLISHERS
NEW YORK, BOSTON, DORDRECHT, LONDON, MOSCOW

eBook ISBN: 0-306-48707-1
Print ISBN: 1-4020-7559-6

©2004 Springer Science + Business Media, Inc.

Print ©2003 Kluwer Academic Publishers
Dordrecht

All rights reserved

No part of this eBook may be reproduced or transmitted in any form or by any means, electronic, mechanical, recording, or otherwise, without written consent from the Publisher

Created in the United States of America

Visit Springer's eBookstore at: http://www.ebooks.kluweronline.com
and the Springer Global Website Online at: http://www.springeronline.com

Table of Contents

Preface vii

Part I: Fractional-N Synthesis

A. van Roermund 1

Practical Design Aspects in Franctional-N Frequency Synthesis
W. Rhee 3

Design and Simulation of Franctional-N Frequency Synthesizers
M. Perrott..27

Monolithic CMOS Fractional-N Frequency Synthesizer Design for High Spectral Purity
B. De Muer, M. Steyaert 51

A 19mW 2.2GHz Fully Integrated CMOS Sigma Delta Fractional Synthesiser with 35Hz Frequency Step and Quantization Noise Compensation
I. Bietti, G. Albasini, E. Temporiti, R. Castello..77

Implementation Aspects of Fractional-N Techniques in Cellular Handsets
Y. Le Guillou, D. Brunel..97

Fractional-N Phase Locked Loops and it's Application in the GSM System
G. Märzinger, B. Neurauter.. 111

Part II: Design for Robustness

M. Steyaert... 129

ESD for Analogue Circuit Design
D. Clarke, A. Righter.. 131

ESD in Smart Power Processes
G. Croce, A. Andreini, L. Cerati, G. Meneghesso, L. Sponton 169

RF-ESD Co-Design for High Performance CMOS LNAs
P. Leroux, M. Steyaert 207

Improvement of System Robustness through EMC Optimazation
B. Deutschmann 227

Robustness in Analog Design
M. De Mey 243

Minimizing Undesired Coupling and Interaction in Mixed Signal ICs
T.J. Schmerbeck 255

Part III: Line and Bus Drivers

J. Huijsing 273

Looking to/for Low Power ADSL Drivers in the DSLAM
E. Moons 275

Class-AB Low-Distortion Drivers for ADSL
T. Ferianz 291

Class D Self-Oscillating Line Drivers
T. Piessens, M. Steyaert 309

Class G/H Line Drivers for xDSL
J. Pierdomenico 333

The USB 2.0 Physical Layer: Standard and Implementation
G. den Besten 359

Backplane Transceivers
K. Tam, W. Ellersick, R. Soenneker 379

PREFACE

This book contains the revised and extended tutorials that 18 experts have presented at the 12th workshop on Advances in Analog Circuit Design (AACD) held at 15-17 April 2003, in Graz, Austria. The book comprises three parts, one per topic, each with 6 tutorial contributions. The three topics are:

- Fractional-N Synthesis;
- Design for Robustness;
- Line and Bus Drivers.

Each topic is introduced with a foreword by the chairman of the day, respectively Arthur van Roermund, Michiel Steyaert, and Han Huijsing. Together, they form the permanent program committee of AACD. The local organising committee of the workshop was formed this year by Herbert Grünbacher (workshop chairman) from Carinthia Tech Institute, Villach, Austria; Wolfgang Pribyl from Austriamicrosystems; and Franz Dielacher from Infineon Technologies.

This book is number 12 in the series called Analog Circuit Design. The topics discussed in previous issues are:

AACD 2002 Spa (Belgium)

- Structured Mixed-Mode Design
- Multi-Bit Sigma-Delta Converters
- Short-Range RF Circuits

AACD 2001 Noordwijk (The Netherlands)

- Scalable Analog Circuits
- High-Speed D/A Converters
- RF Power Amplifiers

AACD 2000 Munich (Germany)

- High-Speed A/D Converters
- Mixed-Signal Design
- PLLs and Synthesizers

AACD 1999 Nice (France)

- XDSL and Other Communication Systems
- RF-MOST Models and Behavioural Modelling
- Integrated Filters and Oscillators

AACD 1998 Copenhagen (Denmark)
- 1-Volt Electronics
- Mixed-Mode Systems
- LNAs and RF Power Amps for Telecom

AACD 1997 Como (Italy)
- RF A/D Converters
- Sensor and Actuator Interfaces
- Low-Noise Oscillators, PLLs and Synthesizers

AACD 1996 Lausanne (Swiss)
- RF CMOS Circuit Design
- Bandpass SD and Other Converters
- Translinear Circuits

AACD 1995 Villach (Austria)
- Low Noise/Power/Voltage
- Mixed-Mode with CAD Tools
- Voltage, Current and Time References

AACD 1994 Eindhoven (The Netherlands)
- Low Power, Low Voltage
- Integrated Filters
- Smart Power

AACD 1993 Leuven (Belgium)
- Mixed-Mode A/D Design
- Sensor Interfaces
- Communication Circuits

AACD 1992 Scheveningen (The Netherlands)
- Opamps
- ADC
- Analog CAD

We hope that also the current book in this series will provide a valuable contribution to our Analog Circuit Design community.

Arthur van Roermund

PART I: FRACTIONAL-N SYNTHESIZERS

Fractional-N synthesis refers to the synthesis of a reference frequency that is related with a fractional ratio to another reference frequency. This is especially important in RF communication electronics, where oscillator frequencies have to be tuned to specific RF frequencies, while coupled to a highly accurate reference frequency, most commonly with a phase-locked loop (PLL). If an integer divider is used in the feedback loop, the reference frequency has to be kept very low to obtain enough frequency resolution, which has drawbacks like strong noise gain and, because of stability reasons, low-bandwidth loops, and thus slow lock-in behaviour. Fractional-N dividers provide a solution here, but if simply done by changing temporarily between two integer values, this results in strong phase perturbations, and therefore in strong undesired phase noise and frequency spurs in the output spectrum. A solution is found in sigma-delta modulation, which can shape the errors in the frequency domain. An optimal trade-off can thus be made between resolution, spectral purity, switching speed, and power dissipation. However, the combination of PLL design, sigma-delta design, and RF circuit design aspects makes this a complex job. Moreover, secondary effects, like noise modulation by non-linear parasitic non-linearities, appear to be significant for high performance integrated RF circuits.

The tutorials that follow all comprise some overview of theory and application of fractional-N synthesis and show experimental verification of it, but from different angles and with different focus. The first paper of Rhee is a general introductory paper including practical design considerations, and a treatment of noise coupling, non-linear effects and out-of-band noise problems. The second paper, of Perrott, focuses on design and simulation techniques and tools, at transfer-function level and detailed behavioural level. The third paper, of De Muer, specifically addresses a fast non-linear analysis method, and shows that due to non-linear effects a single-loop architecture yet has preference over a MASH architecture, and that for high-resolution applications the limitations of fractional-N sigma-delta become apparent. The last three papers are more application oriented. The paper of Bietti focuses on specific designs, down to transistor level, for wireless 3G terminals. The paper of Guillou specifically shows the application in receiver I/Q mixers, UMTS local oscillators, and variable IF in transmitters. The last paper, of Märzinger, addresses the use of fractional-N sigma-delta loops for highly integrated modulators: the fractional-N control input is here driven by the signal that is to be modulated.

Arthur van Roermund

This page intentionally left blank

PRACTICAL DESIGN ASPECTS IN FRACTIONAL-*N* FREQUENCY SYNTHESIS

Woogeun Rhee
IBM Thomas J. Watson Research Center
Yorktown Heights, NY, USA

Abstract

The objective of this work is to present practical design consideration of fractional-*N* frequency synthesis focusing on different aspects from integer-*N* frequency synthesizer design. In this work, a ΔΣ fractional-*N* frequency synthesis technique for high spectral purity is primarily considered. The ΔΣ fractional-*N* synthesizer offers several advantages over integer-*N* synthesizers and conventional fractional-*N* synthesizers. However, superior performance of the ΔΣ fractional-*N* synthesizer can be possibly limited by noise coupling, nonlinearity, and out-of-band phase noise performance. Design efforts and system perspectives both for the ΔΣ modulator and the PLL are needed to overcome those issues. System and circuit design aspects in ΔΣ fractional-*N* synthesizer design have been addressed followed by experimental hardware results.

1. Introduction

The demand for low-cost high-performance frequency synthesizers is growing as wireless systems are diversified. High data-rate systems such as general packet radio service (GPRS) need agile frequency switching and require lower in-band phase noise to maintain the same integrated phase error with wider loop bandwidth. Standard frequency synthesizers based on a PLL have difficulties in meeting various specifications due to the fundamental trade-off between loop bandwidth and channel spacing. The circuit noise floor of the synthesizer and spurious tones can be suppressed by narrowing the PLL bandwidth, but the narrowband PLL suffers from long settling time. The narrow loop bandwidth also

A. van Roermund et al. (eds.), Analog Circuit Design, 3-26.
© 2003 *Kluwer Academic Publishers. Printed in the Netherlands.*

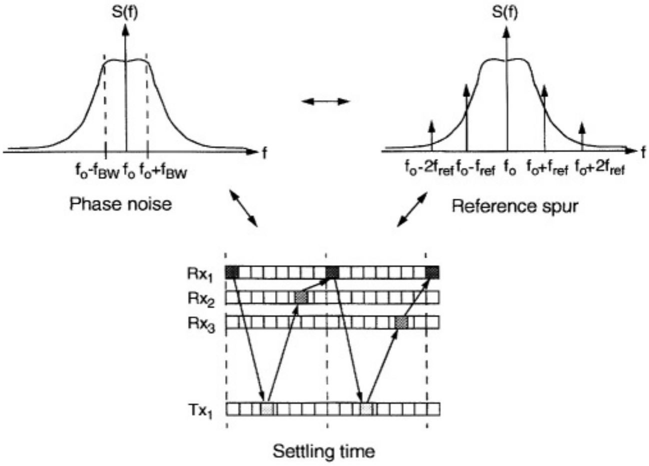

S(f)
f
f_o-f_{BW} f_o f_o+f_{BW}
Phase noise
S(f)
f
f_o-2f_{ref} f_o-f_{ref} f_o f_o+f_{ref} f_o+2f_{ref}
Reference spur
Rx_1
Rx_2
Rx_3
Tx_1
Settling time

Fig. 1. Design trade-off in PLL-based frequency synthesizers.

put stringent requirement for the VCO noise performance. Therefore, there is a trade-off to determine the synthesizer performance in terms of the phase noise, the spurious tone, and the settling time as illustrated in Fig. 1. On the other hand, a fractional-*N* technique provides wide bandwidth with narrow channel spacing and alleviates PLL design constraints for phase noise and reference spur. The inherent problem of the fractional-*N* frequency synthesizer is that periodic operation of the dual-modulus divider produces spurious tones. Among several spur reduction methods, the ΔΣ modulation method is primarily considered in this work.

2. Fractional-*N* Frequency Synthesis

Fractional-*N* frequency synthesizers are PLL-based synthesizers which have a frequency resolution finer than phase detector frequency. The fractional-*N* method originally comes from *Digiphase* technique [1],[2]. Fig. 2 shows the block diagram of the traditional fractional-*N* frequency synthesizer. Fractional division ratio is obtained by periodically modulating the control input of the dual-modulus divider. For example, to achieve an $N + 1/4$ division ratio or the fractional modulo of 4, an $N + 1$ division is done after every three N divisions. The carry of the accumulator is the sequence of {...000100010001...}, where the $N + 1$ division ratio is corresponding to “1”.

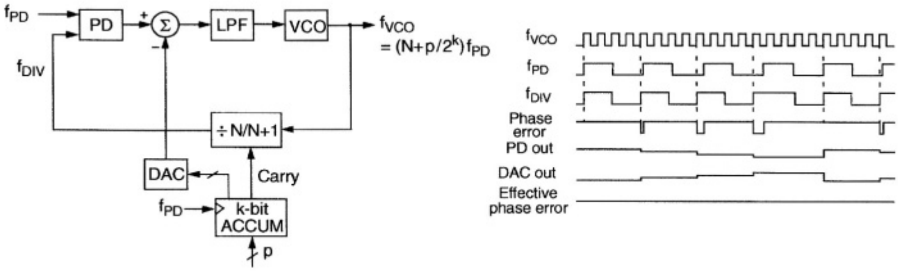


Fig. 2. Fractional-N technique using DAC cancellation.

2.1 Why fractional-*N*?

Since phase detector frequency can be higher than the frequency resolution, the fractional-*N* synthesizers offer several advantages over integer-*N* synthesizers. Firstly, the in-band phase noise contribution from the PLL circuits excluding the VCO is less. For example, to achieve –80 dBc/Hz in-band noise at 2-GHz output with the phase detector frequency of 200 kHz, the PLL circuit noise at the phase detector output should be as low as –160 dBc/Hz due to the multiplication factor of 20log(10 000). When the fractional-*N* method is used with the phase detector frequency of 8 MHz, the phase noise requirement of the PLL circuits becomes only –112 dBc/Hz, which can be easily met even in CMOS. Secondly, the reference spur performance is less sensitive to the leakage current and to the charge pump current mismatch. For example, with a 64-modulo fractional-*N* technique, the leakage current as high as 10 nA and the 10% mismatch of the charge pump output currents may not degrade reference spur performance significantly while they are critical in the integer-*N* frequency synthesizers [3]. Thirdly, the fractional-*N* technique offers agile frequency switching with wide loop bandwidth. Some applications employ narrowband fractional-*N* synthesizers simply to alleviate the PLL noise contribution. In that case, faster settling time can be achieved by using a dynamic bandwidth method. By dynamic bandwidth we mean that the loop bandwidth is set to be wider than the desired one when the PLL is in the frequency acquisition mode [4],[5]. With high phase detector frequency, the loop bandwidth in the transient mode can be set high with less overshoot problem. The fractional-*N* technique provides the opportunity of using dynamic bandwidth methods more effectually.

2.2 Spur reduction techniques

The unique problem of the fractional-N synthesizer is unwanted spur generation which is caused by the periodic operation of the dual-modulus divider. The fractional-N synthesis is not useful in practical applications unless the fractional spurs are suppressed. Thus, additional circuitry must be added to suppress those fractional spurs. Various techniques have been proposed, and their performances are summarized in the literature [6],[7].

The phase interpolation method using a DAC is a traditional spur reduction method employed in digiphase synthesizers. Fig. 2 shows the basic architecture and its operation. Since the phase error is compensated in the voltage domain, this method suffers from analog imperfections. The mismatch results mainly from limited DAC resolution and the limited accuracy of the DAC. This approach is more effective when a sample-and-hold (S/H) phase detector is used. For the S/H phase detector, the DAC needs to match only the dc voltage during one reference clock period. For the PLLs which use a P/FD and a charge pump, the phase interpolation in the time domain instead of in the voltage domain can provide better spur suppression [8].

2.3 ΔΣ modulation in fractional-N frequency synthesis

Fig. 3 shows the fractional-N division scheme using a ΔΣ modulator which achieves fine frequency resolution with a randomly-modulated dual-modulus divider [9],[10]. The ΔΣ modulation technique is similar to the random jittering method [7], but it does not have a $1/f^2$ phase noise spectrum due to noise shaping property. The PLL acts as a low-pass filter to the modulator quantization noise, and the effective oversampling ratio $\mathrm{OSR}_{\mathrm{eff}}$ can be defined by the ratio of the phase detector frequency f_{PD} to the PLL noise bandwidth f_c, or

$$\mathrm{OSR}_{\mathrm{eff}} = \frac{f_{\mathrm{PD}}}{2f_c} \,. \tag{1}$$

Narrowing the PLL noise bandwidth increases the effective oversampling ratio, which results in high in-band signal-to-noise ratio (SNR). When high-order ΔΣ modulators are used, the PLL needs more poles in the loop filter to suppress the quantization noise at high frequencies.

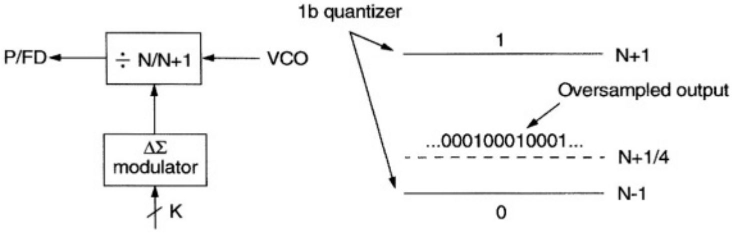


Fig. 3. Basic concept of ΔΣ fractional division.

2.4 ΔΣ modulation vs. conventional digiphase

Conventional digiphase-based fractional-*N* synthesizers suffer from poor fractional spur performance when the analog matching is not well controlled. The instantaneous phase error which needs to be cancelled at the phase detector output is the fraction of the VCO output period, which can be as low as 10^{-3} rad. In the digiphase method, design complexity depends on the VCO output frequency. For applications which do not require agile frequency switching, this method would be useful to relax PLL circuit noise contribution with comparable spur performance. The settling time can be also improved with dynamic bandwidth scheme.

On the other hand, the ΔΣ fractional-*N* synthesizers offer arbitrarily fine frequency resolution regardless of VCO output frequency, which can make the synthesizer compensate for crystal-frequency drift with a digital word and accommodate various crystal frequencies without reducing phase detector frequency [11]-[13]. The all-digital modulation also makes the spur-reduction scheme less sensitive to process variation. With superior phase noise and spur performance, the ΔΣ fractional-*N* synthesis has been extended to a direct digital modulation with a digital baseband filter [14]-[17].

A few-bit ΔΣ modulation is an another interesting way to have moderate performance between the conventional digiphase method and the general ΔΣ modulation method [18]. It features negligible power consumption, small area, and simple design. The method would be useful for narrowband synthesizers to achieve better in-band noise performance.

3. ΔΣ Fractional-*N* Frequency Synthesis: System Design Aspects

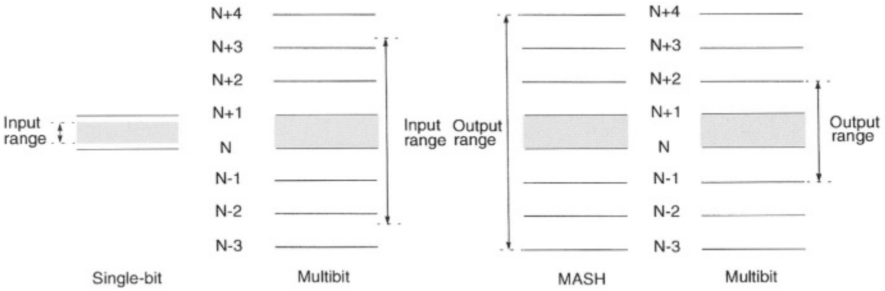
Use of noise-shaped modulators in frequency synthesis alleviates the analog design constraints of the PLL. In this section, a multibit single-loop modulator is presented, and system design aspects over integer-*N* synthesizers are considered.

3.1 Choice of noise transfer function (NTF)

For fractional-*N* frequency synthesis, two types of ΔΣ modulators have been used. One is a single-loop modulator, and the other is a cascaded modulator called MASH. The single-loop modulator has a choice of a single-bit or a multibit output depending on the quantizer while the MASH architecture outputs only multi bits.

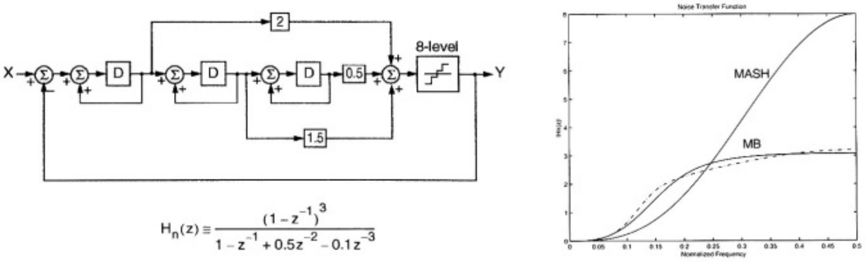
The high-order single-loop ΔΣ modulator with a single-bit quantizer is less sensitive to nonlinearity. The drawback of this architecture is the limited dynamic input range due to stability problem. As shown in Fig. 4(a), the inability to use the full scale of the quantizer makes the frequency synthesizer face the dead-band problem, unless the reference frequency is high enough to cover all channels without changing the integer division ratio. By having a multilevel quantizer, the dynamic input range problem can be solved. The eight-level quantizer in Fig. 4(a) expands the active division range from $\{N, N+1\}$ to $\{N-3, N-2, ..., N+3, N+4\}$ without increasing the minimum quantizer level. Therefore, the multibit high-order modulator can be easily designed to be stable over all interpolated range between N and $N+1$, which is about 12% of the full range of the quantizer.

Compared to the MASH modulator, the multibit single-loop modulator has less high-frequency noise at the phase detector output. Although the MASH topology with the same order can shape the in-band noise more sharply, it produces an output bit pattern spread more widely than the single-loop noise shaper does as shown in Fig. 4(b). Different from the integer-*N* synthesizer, the fractional-*N* synthesizer with a ΔΣ modulator makes the charge pump have the dynamic turn-on time after phase-locked. Less spread output bit pattern makes the synthesizer less sensitive to substrate noise coupling as the modulated turn-on time of the charge pump in the locked condition decreases. It also helps to lessen the nonlinearity problem of the wideband ΔΣ fractional-*N* synthesizer, which will be discussed later.


Input range
N+4
N+3
N+2
N+1
N
N-1
N-2
N-3
Input range
Output range
Single-bit
Multibit
N+4
N+3
N+2
N+1
N
N-1
N-2
N-3
Output range
MASH
Multibit

(a) (b)

Fig. 4. Architecture comparison: (a) single-bit vs. multibit, and (b) MASH vs. multibit.


X
Y
D
D
D
2
0.5
1.5
8-level
+
$H_n(z) = \frac{(1-z^{-1})^3}{1-z^{-1}+0.5z^{-2}-0.1z^{-3}}$
Noise Transfer Function
MASH
MB
Hn(z)
Normalized Frequency
0
1
2
3
4
5
6
7
8
0
0.05
0.1
0.15
0.2
0.25
0.3
0.35
0.4
0.45
0.5

(a) (b)

Fig. 5. (a) Third-order ΔΣ modulator with 3-b quantizer, and (b) NTF.

Fig. 5(a) shows the third-order single-loop modulator with an eight-level quantizer. The NTF is derived from the high-order topology [19]. To avoid digital multiplication, the coefficients of {2, 0.5, 1.5} are implemented by using shift operations. This constraint slightly modifies the original NTF, but it still maintains the causality and the stability conditions [20]. Low-*Q* Butterworth poles are used to reduce the high-frequency shaped noise energy, which results in low spread output bit pattern. As shown in Fig. 5(b), the NTF of the proposed modulator has the passband gain of 3.1 and the corner frequency of 0.18 f_s for the clock frequency f_s. Note that the high corner frequency is preferred for in-band noise suppression, but it increases the high-frequency noise energy [21].

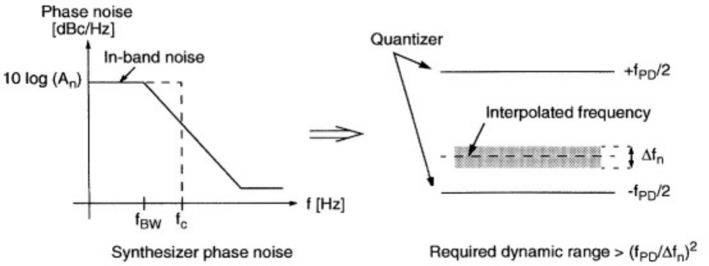

Phase noise [dBc/Hz]
In-band noise
10 log (A$_n$)
f [Hz]
f_{BW}
f_c
Synthesizer phase noise
Quantizer
+f_{PD}/2
Interpolated frequency
Δf_n
-f_{PD}/2
Required dynamic range > (f_{PD}/Δf_n)^2

Fig. 6. Dynamic range consideration in ΔΣ fractional division.

3.2 Dynamic range consideration

By interpreting well-known theoretical results of the oversampling ADC in the frequency domain, we can derive the upper bound of the loop bandwidth in terms of the in-band phase noise, the phase detector frequency, and the order of the ΔΣ modulator. If the in-band phase noise of A_n (rad^2/Hz) of the frequency synthesizer is assumed to be limited within the noise bandwidth of f_c (Hz) as shown in Fig. 6, the integrated frequency noise Δf_n (rms Hz) within f_c is approximately [22]

$$\Delta f_n \cong \sqrt{\frac{2}{3} A_n} \cdot f_c^{\frac{3}{2}} \quad . \tag{2}$$

where $f_c \gg f_o$ assumed. Because the quantizer level in the frequency domain is equivalent to f_{PD} with the frequency noise of Δf_n as illustrated in Fig. 6, the dynamic range of the Lth-order ΔΣ modulator should meet the following condition [23]

$$\frac{3}{2} \cdot \frac{2L+1}{\pi^{2L}} \cdot (\mathrm{OSR}_{\mathrm{eff}})^{2L+1} > \left(\frac{f_{\mathrm{PD}}}{\Delta f_n}\right)^2 \quad , \tag{3}$$

where $\mathrm{OSR}_{\mathrm{eff}}$ is defined in Equation (1). Therefore, from Equations (1), (2), and (3), we obtain

$$f_c < \left[A_n \cdot \frac{L+0.5}{(2\pi)^{2L}}\right]^{\frac{1}{2L-2}} \cdot f_{\mathrm{PD}}^{\frac{2L-1}{2L-2}} \quad . \tag{4}$$

An integrated phase error θ_{rms} [rms rad] is an important factor for synthesizers in digital communications, and it is given by

$$\theta_{\mathrm{rms}} = \sqrt{2A_n \cdot f_c}\ . \tag{5}$$

From Equations (4) and (5), an approximate upper bound of the bandwidth is obtained, or

$$f_c < \left[\left(\frac{\theta_{\mathrm{rms}}}{\sqrt{2}}\right)^2 \cdot \frac{L+0.5}{(2\pi)^{2L}}\right]^{\frac{1}{2L-1}} \cdot f_{\mathrm{PD}}\ . \tag{6}$$

For example, when the phase detector frequency is 8 MHz, the upper bound of the bandwidth with the third-order ΔΣ modulator to meet less than 1°-rms phase error is 195 kHz. In practice, the required loop bandwidth is narrower than that by Equation (6) since the quantization noise of the third-order modulator is tapered off after the fourth pole of the PLL.

3.3 Phase noise

Phase noise is one of the most important specifications for frequency synthesizers. In synthesizer design, it is important to identify phase noise contribution from each source. The noise contributions from various sources for a type-II, fourth-order PLL with a third-order modulator are plotted in Fig. 7. Depending on open-loop gain design, the ΔΣ modulator can affect in-band noise or out-of-band noise. The in-band noise may be limited by PLL nonlinearity. The out-of-band noise can be possibly determined by the residual quantization noise of the modulator rather than the VCO noise. Note that the quantization noise contribution does not depend on division ratio, because it is generated by frequency modulation having the resolution of one VCO clock period. As shown in Fig. 7, the open loop gain needs to be carefully designed to have the overall noise performance meet the system specification. High-order poles are important not only for spur suppression but also for noise performance, which is different from integer-*N* synthesizer design.

Unless the PLL bandwidth is extremely wide, the in-band noise contribution of the ΔΣ modulator is negligible even with the order of two. However, low-order modulators having less uncorrelated output bits may exhibit phase noise

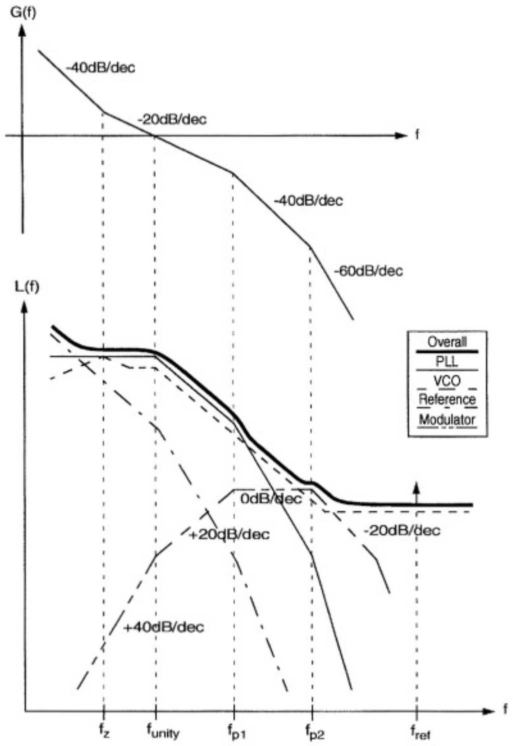


Fig. 7. Overall synthesizer noise including third-order modulator.

fluctuation over time, which will degrade the worst-case phase noise performance. Practically, it is good to have the modulator with the order of at least three in fractional-N synthesizer design.

3.4 Spurious tones

Ideally, second or higher-order $\Delta\Sigma$ fractional-N synthesizers do not create spurious tones. In practice, it is likely to observe fractional spurs at certain output frequencies. Possible mechanisms to generate fractional spurs are listed below.

- Coupling between reference clock and VCO
- PLL nonlinearity
- Idle tones from $\Delta\Sigma$ modulator

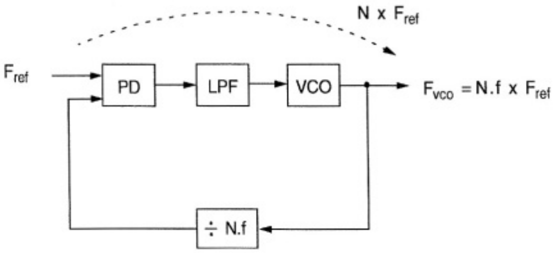


Fig. 8. Fractional spur generation by coupling.

For low-cost synthesizers with limited number of package pins, coupling could be one of major sources to generate fractional spurs. As illustrated in Fig. 8, the harmonics of the reference clock can generate a sideband close to the carrier when the fractional division ratio is close to an integer boundary. Next possibility is PLL nonlinearity, and the detailed description is done in the following section. Idle tones from the $\Delta\Sigma$ modulator itself can be an another possible cause of fractional spur generation. The amount of spur contribution by idle tones is quite low with high-order modulators, but idle tone performance becomes important when PLL nonlinearity is present.

3.5 Settling time

As mentioned previously, the fractional-*N* synthesis offers the opportunity of using the dynamic bandwidth method more effectually. For integer-*N* synthesizers, it is difficult to set much higher charge pump current during the frequency acquisition mode due to stability. With high phase detector frequency, the fractional-*N* synthesizer can accommodate high charge pump current during the transient mode. The dynamic bandwidth scheme is useful to enhance the settling time for the given loop parameter or to have better suppression of spur and out-of-band noise for the given settling time.

3.6 Nonlinearity

In general, the all-digital multibit modulator has no linearity problem, but when it is combined with the PLL, the nonlinearity of the phase detector is a concern. Fig. 9 shows the similarity between the multibit oversampling ADC and the frequency

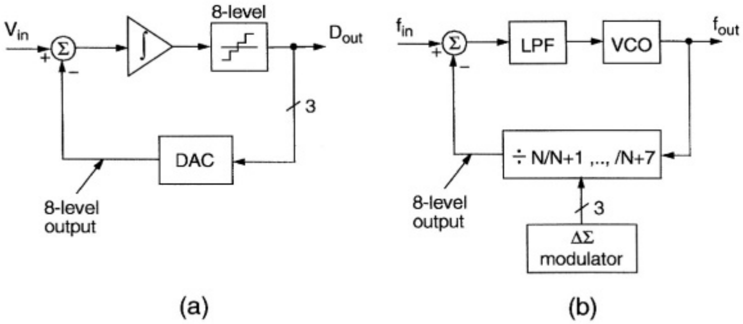


Fig. 9. Multibit oversampling: (a) ADC, and (b) frequency synthesizer.

synthesizer having the multibit modulator. The frequency synthesizer has the multilevel feedback input in the time domain generated by the modulated multimodulus divider, whereas the multibit ADC has the multilevel feedback input in the voltage domain generated by the multibit DAC. It is well known that the multibit DAC performance limits the in-band noise performance as well as the spur performance. Therefore, a similar behavior can be deduced for the multimodulus divider with the modulator. The multimodulus divider conveys the information from the digital domain without having the linearity issue. The phase detector converts the digital quantity into an analog quantity by generating the multiphase errors, and the phase detector nonlinearity is considered a main contributor for nonideal effects of the ΔΣ fractional-*N* synthesizer. In the simulation, the higher-order modulator with nonlinearity still has better spur performance. Therefore, higher-order modulators are preferred not only for lower in-band noise but also for lower spur levels. Since the MASH modulator is based on successive noise cancellation of the previous stage's, the overall performance with nonlinearity may exhibit the characteristic of a first-order modulator regardless of the order of the modulator [20]. For that reason, a 2-1 MASH architecture could be a better choice [24]. A comparison study between the single-loop architecture and the MASH architecture in the nonlinear ΔΣ fractional-*N* synthesizer can be found in the literature [13].

When the PLL bandwidth is set wide, a discrete time *z*-domain model describes loop behavior more accurately than a continuous-time model [25]. One of linearizing assumptions in the *z*-domain analysis is that phase samples occur at constant intervals. The ΔΣ fractional-*N* synthesizer has wide variation of the sampling time in nature, and the amount of relative variation to the reference clock period increases as the division ratio decreases. Therefore, less spread output bit pattern of the single-loop modulator alleviates the nonlinearity problem of the ΔΣ fractional-*N* synthesizer by reducing the nonuniform sampling effect.

Below is the summary of system design aspects in fractional-*N* synthesizer design.

- ΔΣ modulator design needs to be correlated with PLL design
  --> best modulator does not guarantee best fractional-*N* synthesizer!
- Third or higher-order modulator is preferred to reduce idle tone, out-of-band noise, and noise fluctuation effect.
- Fractional spurs might be still there due to coupling, nonlinearity, or modulator idle tones.
- ΔΣ fractional-*N* synthesizer performance is to be ultimately limited by nonlinearity.
- Single-loop ΔΣ modulator gives less spread output bit patterns
  --> useful feature for PLL design to reduce coupling and nonlinearity.

4. ΔΣ Fractional-*N* Frequency Synthesis: Circuit Design Aspects

Fig. 10 shows one example of the dual synthesizer block diagram [12]. The fractional-*N* technique is employed for the RF synthesizer to offer agile channel selection, and a standard integer-*N* architecture is chosen for the IF synthesizer to reduce power, die area, and substrate noise coupling. Implementing a low-cost low-power ΔΣ fractional-*N* synthesizer in a small package is challenging since the modulator induces substantial digital noise. Several circuit design aspects in fractional-*N* synthesizer design are considered in this section.

4.1 Prescaler and frequency divider

Most frequency dividers employ a dual-modulus divider which prescales the division ratio to generate integer division ratio with low-speed frequency counters. With a proper frequency plan, use of the dual-modulus divider helps to save power consumption significantly since only few D flip-flops (DFF) operate at high frequencies. Fig. 11 shows two types of prescalers commonly used in frequency synthesizers. The frequency divider with a $P/P+1$ prescaler has been a

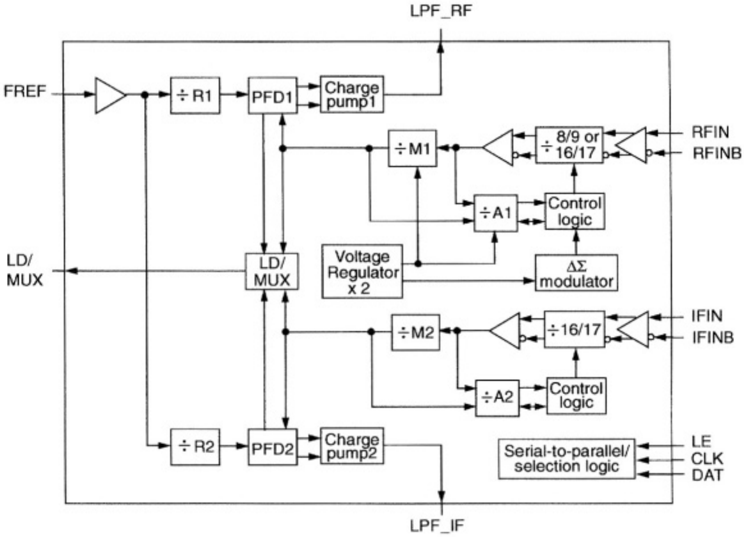
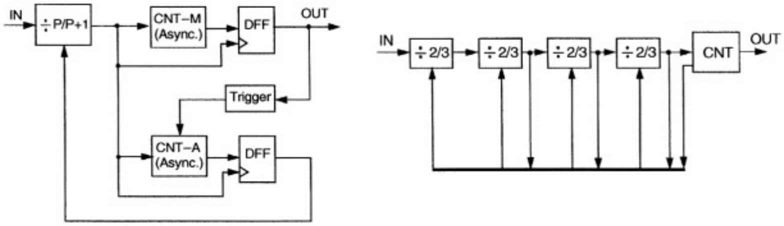

LPF_RF
FREF
÷R1
PFD1
Charge pump1
÷M1
8/9 or ÷16/17
RFIN
RFINB
÷A1
Control logic
LD/ MUX
LD/ MUX
Voltage Regulator x 2
ΔΣ modulator
÷M2
÷16/17
IFIN
IFINB
÷A2
Control logic
÷R2
PFD2
Charge pump2
Serial-to-parallel/ selection logic
LE
CLK
DAT
LPF_IF

Fig. 10. Block diagram example of dual frequency synthesizer [12].

standard topology for most integer-*N* synthesizers, which gives low-power operation with a proper frequency plan. Different from the integer-*N* synthesizer, the fractional-*N* synthesizer has stringent feedback timing requirement for the modulus control as the counters operate at high frequencies. By embedding the DFF at the input of the prescaler modulus control, the timing requirement can be alleviated. Also, having asynchronous counters with the DFF at the output saves power without degrading the performance as shown in Fig. 11(a) [11]. On the other hand, use of 2^k+M frequency prescaler offers lower minimum division ratio [26]. For example, the minimum division ratio of 2^k+M frequency divider with four 2/3 prescalers is only 16 (= 2^4), while the minimum division ratio of frequency divider is 56 (= $8^2 - 8$) with a 8/9 prescaler and 240 (= $16^2 - 16$) with a 16/17 prescaler. Therefore, the 2^k+M frequency divider is useful to accommodate high reference frequency [14]. It generally suffers from harsh modulus control timing for the first stage 2/3 prescaler. Use of 4/5 or 8/9 prescalers in front end will alleviate timing margin when seamless integer division ratio is needed only for certain bands. Use of multiphase prescaler topology can reduce power consumption for both topologies [27], but it requires careful matching among quadrature outputs not to create an additional source of PLL nonlinearity in ΔΣ fractional-*N* synthesizer design.



$N = P \times M + A$

(a)

$N = 2^k + M$ (k=4)

(b)

Fig. 11. (a) Frequency divider: (a) with P/P+1 prescaler, and (b) with 2^k+M prescaler.

4.2 P/FD and charge pump

The fractional-*N* synthesizer is quite immune to phase detector (P/FD and charge pump) mismatches, since phase detector frequency can be much higher than PLL bandwidth. In ΔΣ fractional-*N* synthesizer design, nonlinearity is a concern. Mismatch in the phase detector does not degrade linearity much, but poor phase resolution of the phase detector does. To improve the phase resolution of the phase detector, the switching time of the charge pump needs to be improved. The charge pump which requires less minimum turn-on time (but actual turn-on time set by the P/FD can be longer) to avoid the dead-zone is shown to have better phase resolution. In other words, the rising and falling transitions during current switching must be sharp. Having high phase detector gain for a steep phase-to-voltage transfer function is an another way to improve linearity. It also lowers the impedance of the VCO control line for the given loop bandwidth, which results in reduced coupling toward the VCO input.

4.3 Modulator control logic

Since the input data of the frequency divider are dynamic in the ΔΣ fractional-*N* synthesizer, it is important to synchronize them for proper operation. Also, the dynamic dithering scheme is useful to further decorrelate the modulator outputs [12],[15]. It is important not to reset the ΔΣ modulator logic when the frequency synthesizer is reprogrammed for different channel. Resetting the ΔΣ modulator will impact synthesizer settling time.

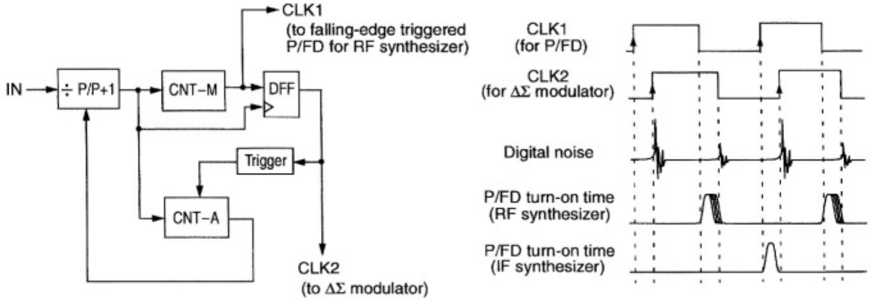

CLK1
(to falling-edge triggered P/FD for RF synthesizer)
IN
÷ P/P+1
CNT–M
DFF
Trigger
CNT–A
CLK2
(to ΔΣ modulator)
CLK1
(for P/FD)
CLK2
(for ΔΣ modulator)
Digital noise
P/FD turn-on time
(RF synthesizer)
P/FD turn-on time
(IF synthesizer)

Fig. 12. Digital noise coupling reduction by timing isolation.

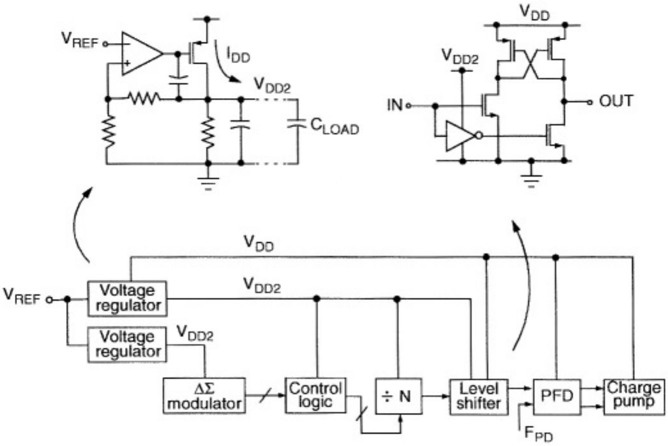

V_REF
I_DD
V_DD2
C_LOAD
V_DD
V_DD2
IN
OUT
V_DD
V_REF
Voltage regulator
V_DD2
Voltage regulator
V_DD2
ΔΣ modulator
Control logic
÷ N
Level shifter
PFD
Charge pump
F_PD

Fig. 13. Digital blocks with on-chip regulator.

4.4 Noise coupling

Dealing with noise coupling is one of difficult tasks for designers. In addition to layout efforts, some ways of reducing noise coupling at circuit level are considered. Fig. 12 shows the timing diagram of the charge pumps and digital blocks in the dual synthesizer. Since the ΔΣ modulator changes states at the rising-edge of the clock, it generates high switching noise at that time. Due to the master-slave DFF operation in the ΔΣ modulator, a significant switching noise is also generated at the falling-edge of the clock. It is important to turn on the charge pump during the quiet period [10],[15]. The IF synthesizer with the narrow loop bandwidth is relatively immune to the high-frequency digital noise. Accordingly, the P/FD in the RF synthesizer is designed to be falling-edge triggered, while the P/FD in the IF synthesizer is rising-edge triggered to minimize the coupling. To further isolate the charge pump turn-on time from the digital switching noise, the clock of the ΔΣ modulator is delayed with a DFF using the high-frequency clock which is available from the prescaler output.

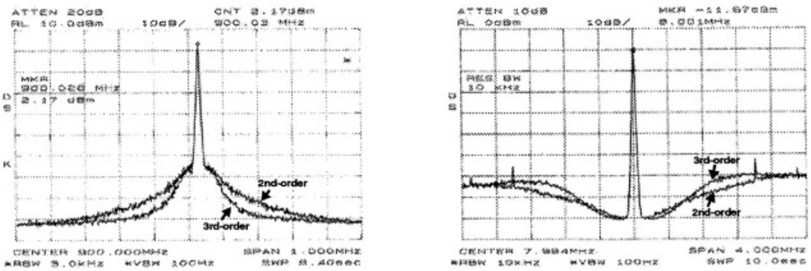
As shown in Fig. 13, an on-chip voltage regulator is designed to have digital blocks operate at lower supply voltage. The voltage regulator reduces the power consumption and the substrate noise coupling from the digital blocks, since lowering the supply voltage of the CMOS digital circuits mitigates the dynamic peak current. For the RF synthesizer, two regulators are implemented; one is used for the modulator and the other for the counters. Miller compensation is done to maintain the stability, and the dc load current I_{DD} enhances the phase margin. A decoupling capacitor is added at the regulator output. Note that the on-chip regulator is used not to improve power supply rejection of the analog circuits but to reduce substrate noise coupling by lowering supply voltage of the digital blocks.

Below is the summary of circuit design aspects in fractional-*N* synthesizer design.

- Frequency divider with $P/P+1$prescaler has lower power, and 2^k+M frequency divider features lower minimum division ratio.
- Linearity is more important than matching in charge pump design --> minimum turn-on time requirement needs to be minimized.
- High charge pump current is still preferred in fractional-*N* synthesizer --> to improve linearity and coupling especially with external VCO.
- Timing isolation and use of lower supply for digital blocks help to reduce noise coupling.

5. Experimental Results

Prototype ΔΣ fractional-*N* synthesizers have been implemented in 0.5-μm CMOS and in 0.35-μm BiCMOS [11],[12]. A type-II fourth-order PLL is designed with a 3-b third-order ΔΣ modulator. Fig. 14(a) shows the measured output spectrum at 900.03 MHz with the 3-b second- and third-order ΔΣ modulators [11]. They are



(a) (b)

Fig. 14. Measured spectra: (a) at VCO output, and (b) at divider output [11].

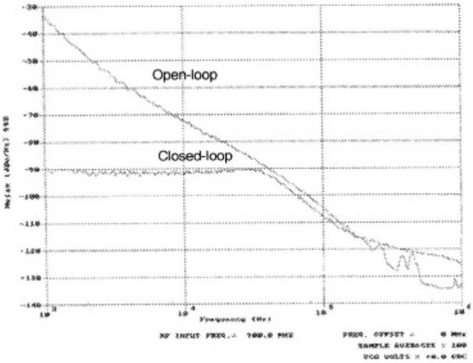
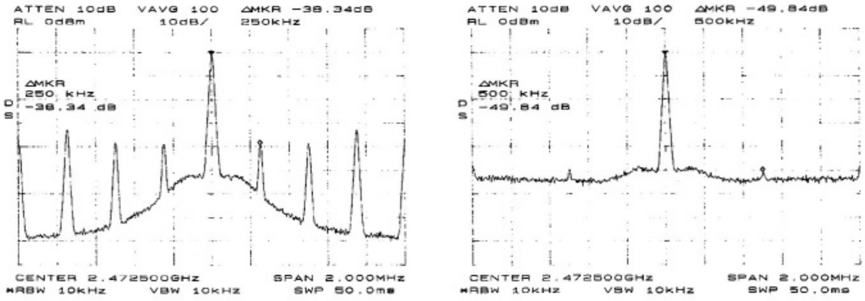


Fig. 15. Measured open-loop and closed-loop output phase noises [11].

compared by switching output bits of each modulator without changing any loop parameter of the synthesizer. The external loop filter is designed to have about 40-kHz loop bandwidth. The third-order modulator case shows less out-of-band noise as expected. Being limited by the synthesizer noise, the output phase noise does not show the noise shaping effect by the $\Delta\Sigma$ modulator unless the loop bandwidth is further widened. However, the effect of the quantization noise from the $\Delta\Sigma$ modulator can be clearly seen at the divider output. Fig. 14(b) shows the shaped quantization noise seen at the divider output. The output with the second-order modulator has idle tones at high frequencies but they can be suppressed to a negligible level at the VCO output with a 40-kHz loop bandwidth. The third-order



(a) (b)

Fig. 16. Measured dithering performance at 2.4725-GHz output: (a) without dithering, and (b) with dithering [12].

modulator does not exhibit high-frequency tones near $f_s/2$. Note that the corner frequency of the quantization noise is close to that of the NTF shown in Fig. 5(b).

Fig. 15 shows the synthesizer output phase noise measured at 900.03 MHz [11]. The phase noise of a free-running external VCO is plotted together. The phase noise floor from 200 to 800 kHz is the residual quantization noise of the modulator. The phase noise is –135 dBc/Hz at 3-MHz offset frequency, and it can be further suppressed either by increasing the phase detector frequency or by pushing high-order poles toward the loop bandwidth sacrificing the phase margin.

Fig. 16 shows the dithering performance in the ΔΣ fractional-*N* synthesizer [12]. To see the dithering effect clearly, only four bits from the MSB are used for the input data of the ΔΣ modulator and the wide loop bandwidth of 150 kHz loop is chosen. With the 8-MHz phase detector frequency, the minimum frequency resolution is 500 kHz, which is the 1/16 of 8 MHz. In Fig. 16, the carrier frequency is set at 2472.5 MHz with the effective division ratio of 309 + 1/16. Figure 16(a) shows the fractional spur performance without dithering. Interestingly, the spur is also generated at 250-kHz offset frequency, which is half of the minimum frequency step. With dithering, the spurs are suppressed and they are smeared into the broadband noise as shown in Fig. 16(b). It is under investigation to clarify whether the spur at 500-kHz offset comes from the idle tone of the modulator with

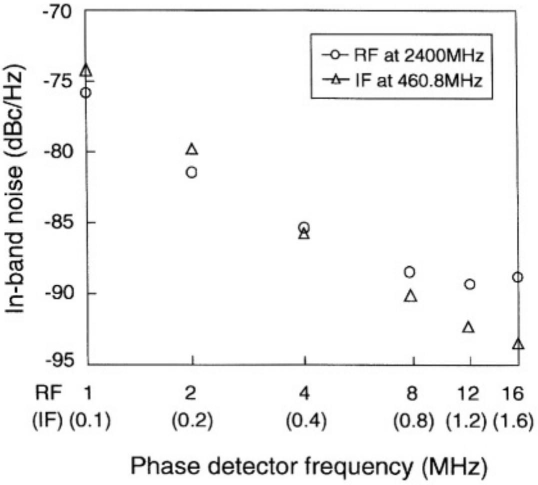


Fig. 17. Measured in-band phase noise vs. phase detector frequency [12].

nonlinearity or the coupling between the harmonic of the reference frequency and the VCO output frequency. In Fig. 16(b), the residual quantization noise of the modulator is substantial since the effective oversampling ratio is only about 50, which is not likely used in practical applications. The in-band noise contribution of the synthesizer with dithering is shown to be –88dBc/Hz at 20-kHz offset frequency, which has about 3-dB worse performance than the one in Fig. 16(a). Fig. 16 implies that the synthesizer with a few bit ΔΣ modulator can be also useful to achieve good phase noise performance with lower power consumption when the system does not require fast settling time for the synthesizer.

Fig. 17 shows the in-band phase noise performance with different phase detector frequencies, which is measured at 2.4-GHz output frequency for the RF synthesizer and at 460.8-MHz output frequency for the IF synthesizer. To make sure that the in-band noise is not limited by the VCO phase noise, the loop bandwidth of 100 kHz is used to suppress the VCO noise to be less than –95 dBc/Hz. The in-band noise performance of the RF synthesizer tends to be saturated at above 8-MHz phase detector frequency, which shows that it is not limited by the synthesizer circuit noise floor, but by either the substrate coupling or the nonlinearity effect of the PLL.

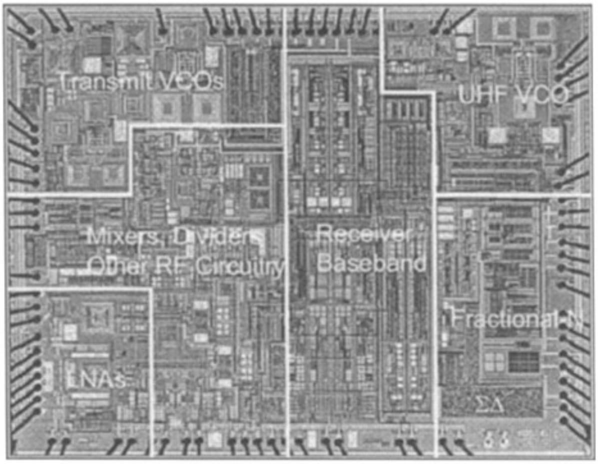


Fig. 18. Fractional-N synthesizer in GPRS direct-conversion transceiver [28].

Another interesting result that is not present in integer-*N* synthesizers is the noise-shaped reference spur, showing that the noise shaping is related with the digital modulation. The noise peaking at the carrier can occur in the under-damped PLL, but the noise shaping at the reference spur is independent of the loop dynamics of the PLL.

Several features which can be observed only in ΔΣ fractional-*N* synthesizers are summarized as follows.

- Arbitrarily fine frequency resolution
- Channel-dependent spur performance
- Noise-shaped reference spur
- "Clear" noise shaping at divider output --> useful feature to check whether multi-modulus divider operates properly with ΔΣ modulator
- Residual out-of-band phase noise floor with low OSR
- Spur and phase noise fluctuation with low-order modulator

Fig. 18 shows the microphotograph of the ΔΣ fractional-*N* synthesizer which is embedded in the quad-band direct conversion GSM/GPRS RF transceiver [28]. To prevent switching noise in the fractional-*N* synthesizer from interfering with the

LNAs, the two circuits are separated by the low-noise baseband analog circuitry. N- and P-type substrate contacts under the baseband supply and ground shunt synthesizer switching noise in the substrate. The degradation of receiver noise figure due to the synthesizer is reported to be less than 0.2 dB [28].

6. Conclusions

Fractional-*N* frequency synthesis technique alleviates several PLL design constraints which are present in integer-*N* synthesizer design. Superior performance of the ΔΣ fractional-*N* synthesizer can be limited by noise coupling, nonlinearity, and out-of-band phase noise performance. Design efforts and system perspectives both for the ΔΣ modulator and the PLL are needed to overcome those issues. In this work, system and circuit design aspects in ΔΣ fractional-*N* synthesizer design have been addressed, and experimental hardware results have been presented for practical understanding of the ΔΣ fractional-*N* synthesizer.

Acknowledgment

The author would like to thank Prof. Bang-Sup Song for his insightful guidance and encouragement throughout this research. The author also gives special thanks to Akbar Ali in Skyworks Solutions (formerly, Conexant Systems) for his strong support and interest in the research of fractional-*N* synthesizers.

References

[1] G. C. Gillette, "The digiphase synthesizer," in *Proceedings of 23rd Annual Frequency Control Symposium,* Apr. 1969, pp. 25-29.

[2] J. Gibbs and R. Temple, "Frequency domain yields its data to phase-locked synthesizer," *Electronics,* pp. 107-113, Apr. 1978.

[3] W. Rhee, "Design of high-performance CMOS charge pumps in phase-locked loops," in *IEEE Proceedings of ISCAS,* vol. 2, June 1999, pp. 545-548.

[4] D. Byrd and C. Davis, "A fast locking scheme for PLL frequency synthesizers," Application Note, National Semiconductor Corporation, July 1995.

[5] B. Memmler, E. Gotz, and G. Schonleber, "New fast-lock PLL for mobile GSM GPRS application," in *Proc. European Solid-State Circuits Conference,* Sep. 2000, pp. 468-471.

[6] D. G. Wilson, W. Rhee, and B. Song, "Integrated RF receiver front ends and frequency synthesizers for wireless," *Emerging Technologies: Designing Low Power Digital Systems, International Symp. on Circuits and Systems,* May 1996.

[7] V. Reinhardt, "Spur reduction techniques in direct digital synthesizers," in *Proceedings of 47th Frequency Control Symposium,* Oct. 1993, pp. 230-241.

[8] W. Rhee and A. Ali, "An on-chip phase-compensation technique in fractional-N frequency synthesis," in *IEEE Proceedings of ISCAS,* vol. 3, May 1999, pp. 363-366.

[9] T. A. Riley, M. Copeland, and T. Kwasniewski, "Delta-sigma modulation in fractional-N frequency synthesis," *IEEE Journal of Solid-State Circuits,* vol. 28, pp. 553-559, May 1993.

[10] B. Miller and R. Conley, "A multiple modulator fractional divider," *IEEE Transactions on Instrumentation and Measurement,* vol. 40, pp. 578-583, June 1991.

[11] W. Rhee, B. Song, and A. Ali, "A 1.1-GHz CMOS fractional-N frequency synthesizer with a 3-b third-order $\Delta-\Sigma$ modulator," *IEEE J. Solid-State Circuits,* vol. 35, pp. 1453-1460, Oct. 2000.

[12] W. Rhee, B. Bisanti, and A. Ali, "An 18-mW 2.5-GHz/900-MHz BiCMOS dual frequency synthesizer with < 10-Hz RF carrier resolution," in *Proc. European Solid-State Circuits Conference,* Sept. 2000, pp. 224-227.

[13] B. De Muer and M. Steyaert, "A CMOS monolithic $\Delta\Sigma$-controlled fractional-N frequency synthesizer for DCS-1800," *IEEE J. Solid-State Circuits,* vol. 37, pp. 835-844, July 2002.

[14] M. Perrott, T. Tewksbury, and C. Sodini, "A 27-mW CMOS fractional-N synthesizer using digital compensation for 2.5-Mb/s GFSK modulation," *IEEE Journal of Solid-State Circuits,* vol. 32, pp. 2048-2060, Dec. 1997.

[15] N. Filiol, T. Riley, C. Plett, and M. Copeland, "An agile ISM band frequency synthesizer with built-in GMSK data modulation," *IEEE Journal of Solid-State Circuits,* vol. 33, pp. 998-1008, July 1998.

[16] S, Willingham, M. Perrott, B. Setterberg, and A. Grzegorek, "An integrated 2.5GHz $\Sigma\Delta$ frequency synthesizer with 5μs settling and 2Mb/s closed loop modulation," in *ISSCC Dig. Tech. Papers,* Feb. 2000, pp. 200-201.

[17] B. Neurauter et al.,"GSM 900/DCS 1800 fractional-*N* modulator with two-point-modulation," in *IEEE MTT-S International Microwave Symp.,* May 2002, pp. 425-428.

[18] R. Ahola and K. Halonen, "A 2GHz ΔΣ fractional-N frequency synthesizer in 0.35 μm CMOS," in *Proc. European Solid-State Circuits Conference,* Sep. 2000, pp. 472-475.

[19] K. Chao, S. Nadeem, W. Lee, and C. Sodini, "A higher-order topology for interpolative modulation for oversampling A/D converters," *IEEE Transactions on Circuits and Systems,* vol. 37, pp. 309-318, Mar. 1990.

[20] S. R. Norsworthy, R. Schreier, and G. C. Temes, *Delta-Sigma Data Converters, Theory, Design, and Simulation.* New York: IEEE Press, 1997.

[21] P. Ju and D. Vallancourt, "Quantization noise reduction in multibit oversampling Σ–Δ A/D convertors," *Electronics Letters,* vol. 28, pp. 1162-1163, June 1992.

[22] Kamilo Feher et al., *Telecommunications Measurements, Analysis, and Instrumentation.* Englewood Cliffs, NJ: Prentice Hall, 1987.

[23] J. C. Candy and G. C. Temes, *Oversampling Delta-Sigma Data Converters.* New York: IEEE Press, 1992.

[24] C. H. Heng and B. Song, "A 1.8GHz CMOS fractional-N frequency synthesizer with randomized multi-phase VCO," in *Proc. Custom Integrated Circuits Conference,* May 2002, pp. 427-430.

[25] J. P. Hein and J. W. Scott, "z-domain model for discrete-time PLL's," *IEEE Transactions on Circuits and Systems,* vol. 35, pp. 1393-1400, Nov. 1988.

[26] Noel E. Hogue, "High speed programmable prescaler," *U.S. patents 4,575,867,* Mar. 1986.

[27] J. Craninckx and M. Steyaert, "A 1.75-GHz/3-V dual modulus divide-by-128/129 prescaler in 0.7 μm CMOS," *IEEE Journal of Solid-State Circuits,* vol. 31, pp. 890-897, July 1996.

[28] R. Magoon, A. Molnar, J. Zachan, G. Hatcher, and W. Rhee, "A single-chip quad-band (850/900/1800/1900 MHz) direct conversion GSM/GPRS RF transceiver with integrated VCOs and fractional-*N* synthesizer," *IEEE Journal of Solid-State Circuits,* vol. 37, pp. 1710-1720, Dec. 2002.

DESIGN AND SIMULATION OF FRACTIONAL-N FREQUENCY SYNTHESIZERS

Michael H. Perrott
Massachusetts Institute of Technology
http://www-mtl.mit.edu/~perrott

Copyright © 2003 by Michael H. Perrott
All rights reserved.

Abstract

Design and simulation techniques are presented for fractional-N frequency synthesizers that allow fast and straightforward assessment of these systems at the transfer function and behavior levels, respectively. The design approach uses a freely available software tool to perform the loop filter design and allow assessment of the impact of parasitic poles/zeros, gain and pole/zero variations, and detector and VCO noise. The simulation techniques, which are also embedded in a freely available software tool, allow fast and accurate simulation of fractional-N synthesizers by leveraging an area conservation approach and by combining the VCO and divider blocks. The simulation approach is verified by comparison to measured results from a custom fractional-N frequency synthesizer IC.

1. Introduction

Fractional-N frequency synthesizers provide high speed frequency sources that can be accurately set with very high resolution, which is of significant value to many communication systems. Figure 1 illustrates this PLL architecture, which consists of a phase-frequency detector (PFD), charge pump, loop filter, voltage controlled oscillator (VCO), and a frequency divider that is dithered between integer values to achieve fractional divide ratios. The realization of fractional divide ratios allows the synthesizer to achieve very high frequency resolution. In this paper, we will focus on a particular class of fractional-N frequency synthesizers for which the divide value is dithered according to the output of a Σ–Δ modulator [1,9].

A. van Roermund et al. (eds.), Analog Circuit Design, 27-49.
© 2003 *Kluwer Academic Publishers. Printed in the Netherlands.*

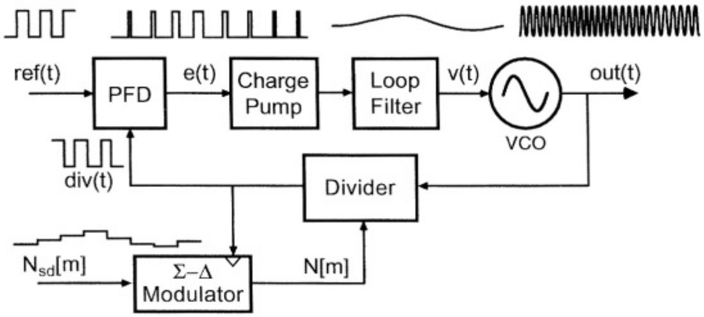

ref(t)
PFD
e(t)
Charge Pump
Loop Filter
v(t)
out(t)
VCO
div(t)
Divider
$N_{sd}[m]$
Σ–Δ Modulator
N[m]

Fig.1. A Σ–Δ Fractional-N Frequency Synthesizer.

Dithering of the divide value by the Σ–Δ modulator allows high frequency resolution to be achieved, but also has the negative side effect of introducing quantization noise that degrades the overall PLL noise performance [1]. It is highly desirable to be able to calculate and simulate the impact of this quantization noise, along with other noise sources in the PLL shown in Figure 2, on the overall PLL performance. It is also desirable to calculate and simulate the dynamic response of the synthesizer in response to variations of the Σ–Δ input in order to evaluate stability and characterize the performance of the system when it is used as a transmitter [2].

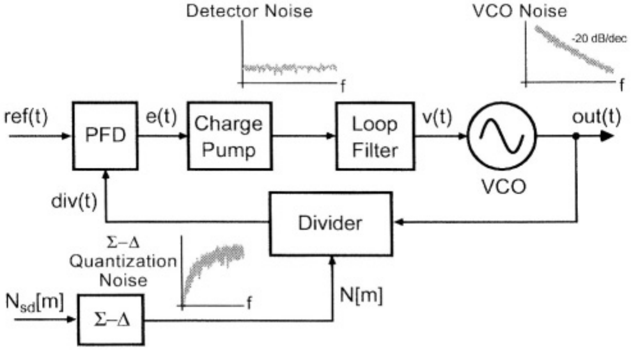

Detector Noise
f
VCO Noise
-20 dB/dec
f
ref(t)
PFD
e(t)
Charge Pump
Loop Filter
v(t)
out(t)
VCO
div(t)
Divider
Σ–Δ Quantization Noise
f
$N_{sd}[m]$
Σ–Δ
N[m]

Fig.2. Noise Sources in a Σ–Δ Fractional-N Frequency Synthesizer.

In this paper we will introduce design and simulation techniques for Σ–Δ fractional-N frequency synthesizers that focus on both dynamic and noise performance. The design techniques are embedded in a GUI-based program called The PLL Design Assistant program, which enables assessment of the synthesizer noise and dynamic performance based on the modeling approach described in [3]. The simulation techniques are embedded in a C++ simulator called CppSim, which enables verification of the synthesizer noise and dynamic performance at a behavioral level. Each of these tools can be freely downloaded from **http://www-mtl.mit.edu/~perrott.** We will first focus on the design techniques, and then on key aspects of the simulation method.

2. Synthesizer Design using the PLL Design Assistant Program

The PLL Design Assistant program allows fast and straightforward design of phase locked loops at the transfer function level. In particular, the program takes as input a desired closed loop transfer function description and then automatically calculates the open loop parameters that must be chosen to achieve the design. The resulting closed loop pole/zero locations, transfer function, and step response are plotted with a simple touch of a button. The impact of non-idealities such as open loop gain and open loop pole variations on the closed loop response can be explored by entering in the variation values into the tool and observing the resulting shifts in the closed loop pole/zero locations, transfer function, and step response. In addition, the impact of parasitic poles and zeros can be accounted for by entering them into the tool and observing the resulting closed loop response. Finally, an estimate of the PLL noise performance can be viewed by entering in noise parameters such as the magnitude of detector noise and VCO noise and observing the resulting phase noise and rms jitter at the PLL output. In this chapter, we will focus only on calculation of the open loop parameters given a desired closed loop specification – the other features of the software are described in detail in its accompanying manual at the author's website indicated above.

2.1. Background

A linearized model of the Σ–Δ fractional-N frequency synthesizer presented in [3] is shown in Figure 4. Here we are assuming that an XOR-based [8] or Tristate [4] phase-frequency detector (PFD) is used, which is then fed into a charge pump with current I_{cp} (please see [3] for more details). As discussed in [3], this linearized model can be used to analyze the "small signal" dynamic properties of the PLL as well as its noise performance. By "small signal", we consider only variations in the PLL frequency caused by small changes in the divide value. If the divide value is stepped by a large value, the PLL will cycle slip many times before re-acquiring frequency lock, and thereby invalidate our modeling assumptions. Fortunately, a well designed Σ–Δ fractional-N frequency synthesizer system can always be ramped in frequency slowly enough to avoid losing frequency lock, so that our modeling assumptions will be accurate for most practical cases. Otherwise, if an estimate of the nonlinear behavior exhibited during cycle slipping is desired, simulation tools such as described later in this chapter can be used.

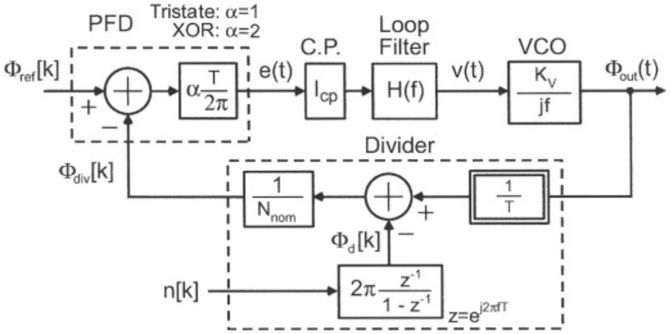
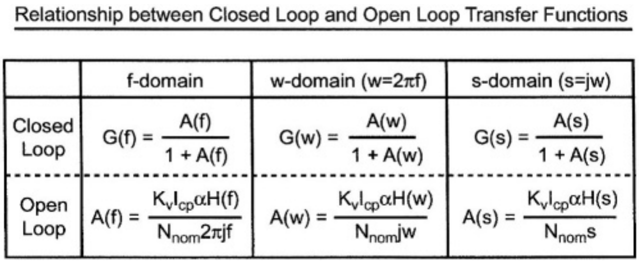


Fig.4. Linearized model of the Σ–Δ fractional-N frequency synthesizer.

The model shown in Figure 4 can be used to design the PLL dynamics to be stable using classical techniques based on examining the phase margin of its open loop response. However, the PLL Design Assistant software allows a much more direct means of design by enabling the user to directly set the characteristics of the closed loop dynamics rather than inferring the closed loop behavior from open loop analysis. The key to the approach is to first design the closed loop dynamics according to desired specifications, and then calculate the open loop response required to realize the desired closed loop behavior,

As discussed in [3], the closed loop response of the system can be parameterized in terms of a single transfer function we label as G(f). Table 1 describes the relationship between G(f) and its open loop counterpart, A(f). The open loop transfer function is, in turn, calculated from Figure 4 as specified in the table. Note that, throughout this document, we will switch between f, w, and s as the frequency variables of interest. In every case, we will assume that $w=2\pi f$ and $s=jw$.



Relationship between Closed Loop and Open Loop Transfer Functions

	f-domain	w-domain (w=2πf)	s-domain (s=jw)
Closed Loop	$G(f) = \frac{A(f)}{1 + A(f)}$	$G(w) = \frac{A(w)}{1 + A(w)}$	$G(s) = \frac{A(s)}{1 + A(s)}$
Open Loop	$A(f) = \frac{K_v I_{cp} \alpha H(f)}{N_{nom} 2\pi jf}$	$A(w) = \frac{K_v I_{cp} \alpha H(w)}{N_{nom} jw}$	$A(s) = \frac{K_v I_{cp} \alpha H(s)}{N_{nom} s}$

Table.1. Relationship between G(f) and A(f).

Once the open loop transfer function is calculated, the last step in the design process is to realize an appropriate loop filter implementation. The methodology of doing so will be explained later in this document.

2.2. Definitions

Given the above background, we can summarize the design process as

1) Designing G(f) to achieve desired closed loop characteristics,
2) Designing A(f) to realize the desired G(f).

In this section, we will outline notation and background to provide understanding of how to accomplish step 1. The following section will provide details describing how step 2 is to be achieved.

2.2.1. Bandwidth, Order, and Shape

Figure 5 illustrates the definitions of bandwidth and order for G(f) that will be assumed in this document.

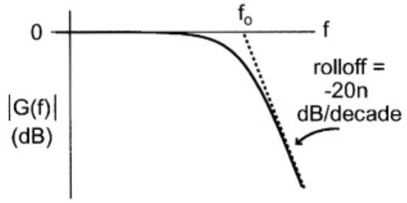


Fig.5. Definition of order and bandwidth for G(f).

Bandwidth is denoted as f_o, and is defined in an asymptotic manner as depicted in the figure above. This definition of bandwidth turns out to be more straightforward for calculations than the traditional 3 dB bandwidth specification. As an example, the asymptotic bandwidth of

$$G(f) = \frac{1}{(1+s/(2\pi f_p))^2}$$

is simply f_p, while the 3 dB bandwidth is $.6423*f_p$.

Order is denoted as n, and is defined according to the rolloff characteristic of the magnitude of G(f) rather than the number of independent state variables in the system. The reason for this definition is that the rolloff of G(f) is more directly of interest to the designer than the number of independent state variables.

The shape of G(f) is defined according to standard filter design methodologies. In particular, it describes the shape of the transfer function and corresponding step response for a given specification of order and bandwidth. Typical examples of shape specifications include Butterworth, Bessel, and Chebyshev.

It is important to note that the above definition of shape considers only the dominant poles in the PLL system. Loosely speaking, the dominant poles are the ones that most strongly influence the behavior of G(f) and are directly specified to achieve the desired G(f) characteristic. In practice, any practical system also includes parasitic poles and zeros that have second order effects on G(f). Discussion of the impact of such parasitics is deferred to the design manual that comes with the PLL Design Assistant program.

2.2.2. Type

The type of a PLL is defined as the number of integrators in its open loop transfer function. In practical PLL implementations, the type is either I or II (note that the VCO contributes one integrator, so the type must be at least I). A type I PLL does not contain an integrator in its loop filter transfer function, whereas a type II PLL does contain an integrator in its loop filter transfer function. Most PLL's are implemented as type II systems, but type I systems have an advantage of having faster settling times [6].

The primary advantage of including an integrator in the loop filter transfer function, as encountered in type II systems, is that it allows the output of the loop filter to achieve an arbitrary DC value while simultaneously forcing the phase error (i.e., the input of the loop filter) to have a steady-state value of zero. Forcing the phase error to zero is quite advantageous in many applications – the phase detector maintains a consistent DC operating point, and the value of the phase error is well controlled.

2.3. Computation of G(f)

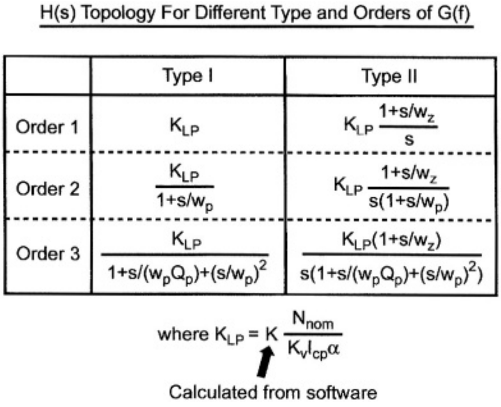
Given the definitions in the previous section, the transfer function G(f) is designed automatically by the PLL Design Assistant software package. Specifically, the user simply enters in desired values for bandwidth, order, shape, type, and f_z/f_o, and the corresponding G(f) transfer function is automatically computed using classical filter design routines. The next section describes how to design the open loop PLL transfer function to achieve the desired G(f) response.

2.4. Loop Filter Design

The key PLL component used to achieve a desired PLL transfer function is the loop filter. In particular, the loop filter implements desired poles and zeros for realizing a given open loop transfer function and, in combination with the charge pump, also sets the overall open loop gain of the system. Three steps must be undertaken in designing the loop filter – selection of its transfer function, selection of its topology, and selection of its transfer function values (and, thereby, selection of the component values in its chosen topology).

2.4.1. Transfer Function Selection

To design the loop filter to achieve a desired G(f) specification, both its transfer function and the value of parameters associated with that transfer function must be properly chosen. The appropriate loop filter transfer function required to realize a given G(f) function is found using the lookup table shown in Table 2. The relevant specifications for this task are the desired order and type of G(f) – note that the bandwidth and shape of G(f) are controlled by the parameter values of the given loop filter transfer function. It should be noted that the table only includes a limited set of orders (1, 2, and 3) and types (I and II), but these are sufficient for most PLL applications due to the prohibitively high analog complexity required to achieve higher order or higher type values.



H(s) Topology For Different Type and Orders of G(f)

	Type I	Type II
Order 1	K_{LP}	$K_{LP}\frac{1+s/w_z}{s}$
Order 2	$\frac{K_{LP}}{1+s/w_p}$	$K_{LP}\frac{1+s/w_z}{s(1+s/w_p)}$
Order 3	$\frac{K_{LP}}{1+s/(w_pQ_p)+(s/w_p)^2}$	$\frac{K_{LP}(1+s/w_z)}{s(1+s/(w_pQ_p)+(s/w_p)^2)}$

where $K_{LP} = K\frac{N_{nom}}{K_v I_{cp}\alpha}$

Calculated from software

Table.2. Loop filter transfer function for G(f) of different order and types.

Examination of Table 2, as well as Figure 4, reveals that the desired DC gain of the loop filter, K_{LP}, is a function of a gain term, K, and the PLL parameters:

1) Nominal divide value, N_{nom},
2) VCO gain, K_v (in units of Hz/V),
3) PFD topology, α,
4) Charge pump current, I_{cp}.

The gain term, K, will be determined automatically for a given G(f) specification by the PLL Design Assistant software. As for the other parameters, the nominal divide value is set as the ratio of the output frequency to the choice of reference frequency, the VCO gain is typically set by the desired frequency range of the PLL, and the PFD topology is set according to the benefits of an XOR-based versus Tristate design. However, the charge pump current is a free parameter, and can be adjusted to either increase or decrease the required value of K_{LP}. The appropriate setting of the charge pump current is determined by the required noise performance of the PLL – a higher charge pump current value yields lower detector noise in general. The maximum value of the charge pump current is typically constrained by limits on the loop filter component values or power dissipation.

2.4.2. Circuit Topology Selection

Given that the loop filter transfer function has been chosen according to Table 2, an appropriate loop filter topology is then selected to realize that transfer function. Popular choices for loop filter topologies can be divided into two categories: passive and active.

Figure 10 illustrates possible loop filter implementations using passive components. The advantage of the passive approach is that good noise performance can be achieved with minimal power consumption, and that the overall PLL implementation is simplified by minimizing the amount of active circuitry. These advantages have led to widespread use of the RC network shown for the Type II, Order 2 implementation in Figure 10. Passive approaches are rarely used for higher order loop filter implementations due to the awkwardness of using inductors.

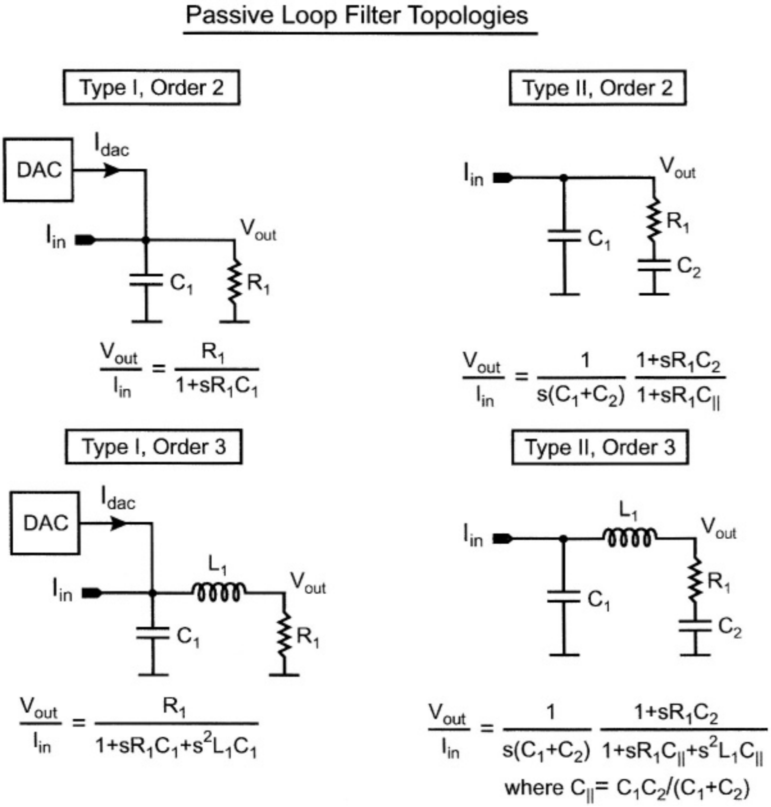

Passive Loop Filter Topologies
Type I, Order 2
DAC
I_{dac}
I_{in}
V_{out}
C_1
R_1
$\frac{V_{out}}{I_{in}} = \frac{R_1}{1+sR_1C_1}$
Type II, Order 2
I_{in}
V_{out}
C_1
R_1
C_2
$\frac{V_{out}}{I_{in}} = \frac{1}{s(C_1+C_2)} \frac{1+sR_1C_2}{1+sR_1C_{||}}$
Type I, Order 3
DAC
I_{dac}
I_{in}
L_1
V_{out}
C_1
R_1
$\frac{V_{out}}{I_{in}} = \frac{R_1}{1+sR_1C_1+s^2L_1C_1}$
Type II, Order 3
I_{in}
L_1
V_{out}
C_1
R_1
C_2
$\frac{V_{out}}{I_{in}} = \frac{1}{s(C_1+C_2)} \frac{1+sR_1C_2}{1+sR_1C_{||}+s^2L_1C_{||}}$
where $C_{||} = C_1C_2/(C_1+C_2)$

Fig.10. Passive loop filter topologies.

Figure 11 illustrates active counterparts to the passive topologies shown in Figure 10. There are two advantages to using an active, rather than passive, approach for the loop filter implementation: 1) the charge pump need not support a wide voltage range at its output since the active approach consistently sets its value to that of a reference voltage, 2) third order PLL dynamics can be achieved without requiring inductors.

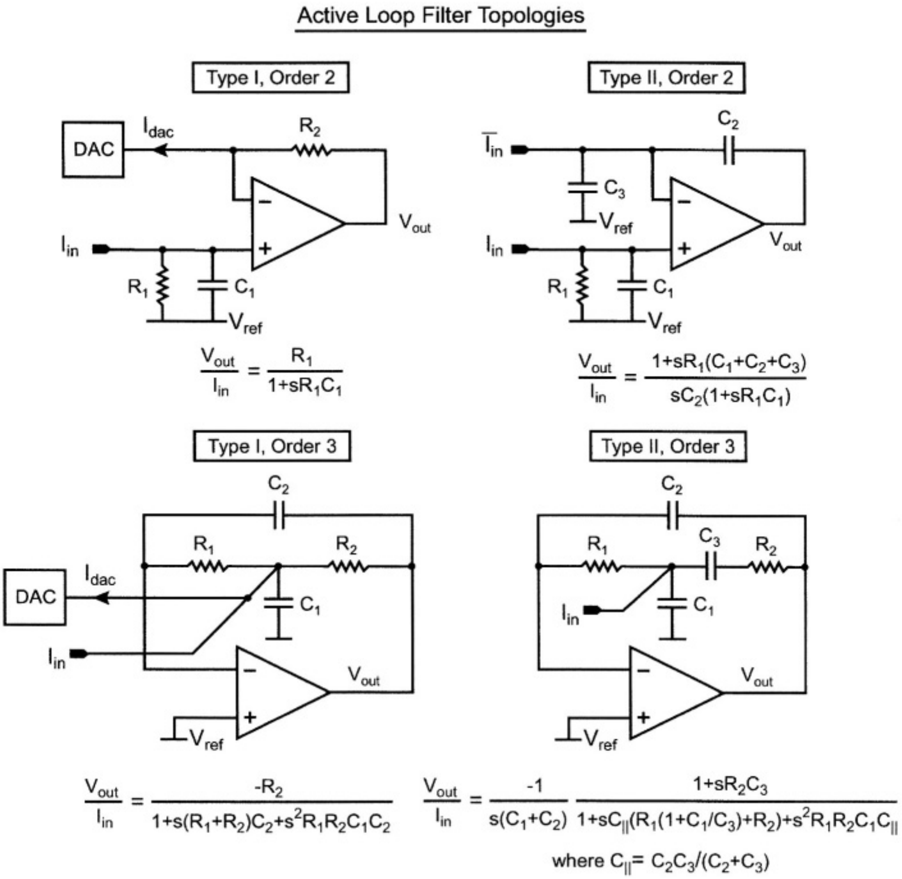

Active Loop Filter Topologies
Type I, Order 2
DAC
I_{dac}
R_2
V_{out}
I_{in}
R_1
C_1
V_{ref}
$\frac{V_{out}}{I_{in}} = \frac{R_1}{1+sR_1C_1}$
Type II, Order 2
C_2
$\overline{I}_{in}$
C_3
V_{ref}
I_{in}
R_1
C_1
V_{ref}
V_{out}
$\frac{V_{out}}{I_{in}} = \frac{1+sR_1(C_1+C_2+C_3)}{sC_2(1+sR_1C_1)}$
Type I, Order 3
C_2
R_1
R_2
DAC
I_{dac}
C_1
I_{in}
V_{out}
V_{ref}
$\frac{V_{out}}{I_{in}} = \frac{-R_2}{1+s(R_1+R_2)C_2+s^2R_1R_2C_1C_2}$
Type II, Order 3
C_2
R_1
C_3
R_2
I_{in}
C_1
V_{out}
V_{ref}
$\frac{V_{out}}{I_{in}} = \frac{-1}{s(C_1+C_2)}\ \frac{1+sR_2C_3}{1+sC_{||}(R_1(1+C_1/C_3)+R_2)+s^2R_1R_2C_1C_{||}}$
where $C_{||} = C_2C_3/(C_2+C_3)$

Fig.11. Active loop filter topologies.

In general, there are two design rules that should be followed when using an active loop filter implementation. First, the output of the charge pump should always feed directly into a high-Q capacitor (i.e. a capacitor with minimal series resistance) in order to attenuate its high frequency content before it feeds into the loop filter opamp. The reason for doing so is that the opamp has limited bandwidth and can exhibit nonlinear behavior if it is directly driven with the high frequencies that are present in the charge pump output. The second rule is that the feedback of the opamp should be configured to achieve unity gain from the opamp terminals to its output. By doing so, the input referred noise of the opamp is not amplified in its influence on the loop filter output. The active topologies shown in Figure 11 achieve both of these desired characteristics.

2.4.3. Computation of Parameters

Given that the loop filter transfer function and its corresponding topology have been selected, the next step is to calculate the parameters of the transfer function and then the component values of its corresponding topology. It is at this point that we now turn to the PLL Design Assistant software as it will perform this task for us. We will use an example to illustrate the relevant principles.

2.4.4. Example Design

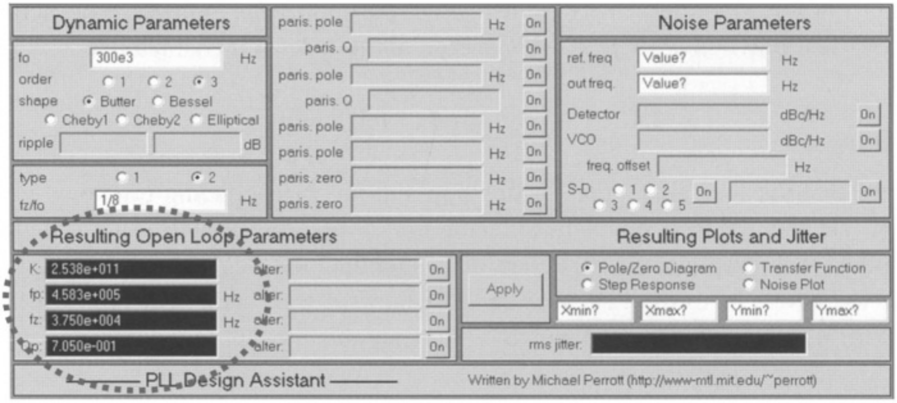
Design a third order, Type II PLL with a Butterworth filter response and bandwidth of f_o=300 kHz. Choose the ratio of f_z/f_o to be 1/8.

Step 1: Select the loop filter transfer function based on Table 2:

$$H(s) = \frac{K_{LP}(1+s/w_z)}{s(1+s/(w_pQ_p)+(s/w_p)^2)} \quad \text{where } K_{LP} = K\frac{N_{nom}}{K_v I_{cp}\alpha}$$

Step 2: Select the corresponding loop filter topology.

Step 3: Solve for K, w_p, Q_p, and w_z in H(s) by using the software tool. Once these values are known, they are used to solve for the component values in the chosen loop filter topology. Figure 12 illustrates the entry of the desired G(f) parameters into the tool along with the resulting loop filter parameter values that it calculates (note that $w_p = 2\pi f_p$ and $w_z = 2\pi f_z$). Figure 13 displays the resulting closed loop step response and pole/zero locations of the PLL. Note that one can also plot the corresponding closed loop transfer function if desired. In addition, the user can set the axis limits for all plots in the designated boxes, and can zoom into a portion of the plot figure by simply dragging a box across the desired plot area with the mouse.



Dynamic Parameters

fo 300e3 Hz

order ○ 1 ○ 2 ● 3

shape ● Butter ○ Bessel ○ Cheby1 ○ Cheby2 ○ Elliptical

ripple dB

type ○ 1 ● 2

fz/fo 1/8 Hz

Resulting Open Loop Parameters

K: 2.538e+011

fp: 4.583e+005 Hz

fz: 3.750e+004 Hz

Qp: 7.050e-001

Noise Parameters

ref. freq Value? Hz

out freq Value? Hz

Resulting Plots and Jitter

● Pole/Zero Diagram ○ Transfer Function ○ Step Response ○ Noise Plot

Xmin? Xmax? Ymin? Ymax?

PLL Design Assistant

Written by Michael Perrott (http://www-mtl.mit.edu/~perrott)

paris. pole Hz On
paris. Q On
paris. pole Hz On
paris. Q On
paris. pole Hz On
paris. pole Hz On
paris. zero Hz On
paris. zero Hz On
Detector dBc/Hz On
VCO dBc/Hz On
freq. offset Hz
S-D ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 On On
alter On
alter On
alter On
alter On
Apply
rms jitter

Fig.12. Calculate loop filter parameter values using the PLL Design Assistant.

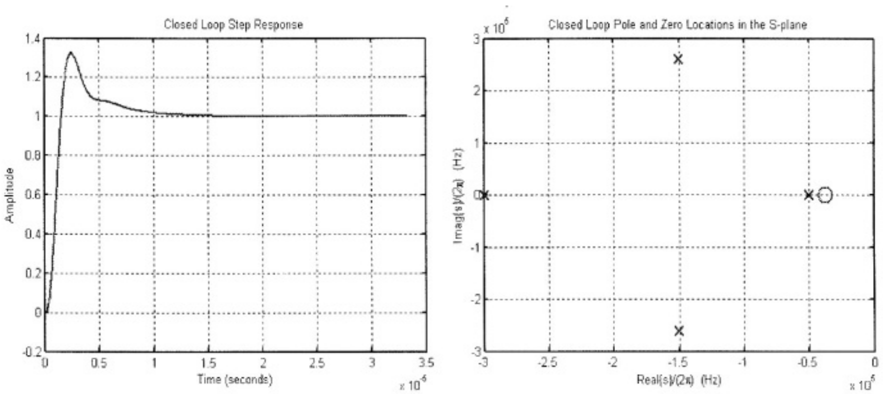


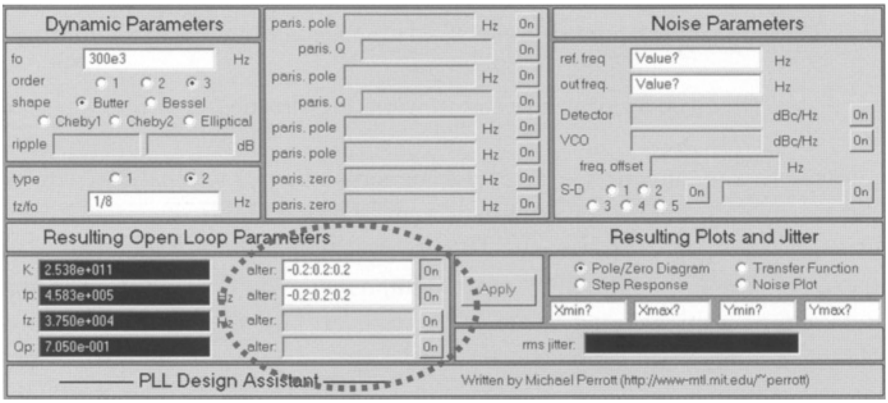
Fig.13. Plot of closed loop step response and pole/zero locations using the tool.

2.5. Impact of Open Loop Parameter Variations

In a practical PLL implementation, the value of the open loop gain and loop filter pole and zero locations will vary from chip to chip and across temperature. Classical PLL design methodologies address this issue by examining the phase margin of the open loop transfer function across all such variations to infer whether the system remains stable. The problem with this approach is that it provides little information on other important issues such as the change in settling time or closed loop transfer function that will occur due to such changes.

Using the PLL Design Assistant, one can directly examine the changes in the closed loop step response, frequency response, and pole/zero locations by simply entering in the variations of each open loop parameter into the tool.

Figure 14 displays the inclusion of open loop parameter variations in the tool. These variations are entered using Matlab notation, so that, in this case, the open loop gain and loop filter pole values are varied by percentage changes of {-20%, -10%, 0, 10%, 20%} as specified by the notation -0.2:0.2:0.2.



Dynamic Parameters
fo: 300e3 Hz
order: ○ 1 ○ 2 ● 3
shape: ● Butter ○ Bessel ○ Cheby1 ○ Cheby2 ○ Elliptical
type: ○ 1 ● 2
fz/fo: 1/8 Hz

Noise Parameters
ref. freq: Value? Hz
out freq.: Value? Hz

Resulting Open Loop Parameters
K: 2.538e+011 — alter: -0.2:0.2:0.2
fp: 4.583e+005 Hz — alter: -0.2:0.2:0.2
fz: 3.750e+004 Hz — alter:
Qp: 7.050e-001 — alter:

Resulting Plots and Jitter
● Pole/Zero Diagram ○ Transfer Function ○ Step Response ○ Noise Plot

paris. pole
paris. Q
paris. pole
paris. Q
paris. pole
paris. pole
paris. zero
paris. zero
Hz
On
ripple
dB
Detector
dBc/Hz
VCO
dBc/Hz
freq. offset
S-D ○ 1 ○ 2 ○ 3 ○ 4 ○ 5
Apply
Xmin?
Xmax?
Ymin?
Ymax?
rms jitter:
PLL Design Assistant
Written by Michael Perrott (http://www-mtl.mit.edu/~perrott)

Fig.14. Examine the impact of changes in the open loop parameters.

Figure 15 displays the resulting closed loop step response and pole/zero locations due to the above open loop parameter variations. The step response plots allow us to quickly assess that the system has no stability issues over the range of parameter variations entered. The pole/zero locations also confirm that the system remains stable as evidenced by the fact that the poles remain in the left half of the S-plane. Note that the user can effectively do root locus plots as a function of open loop gain or other open loop parameters with this feature.

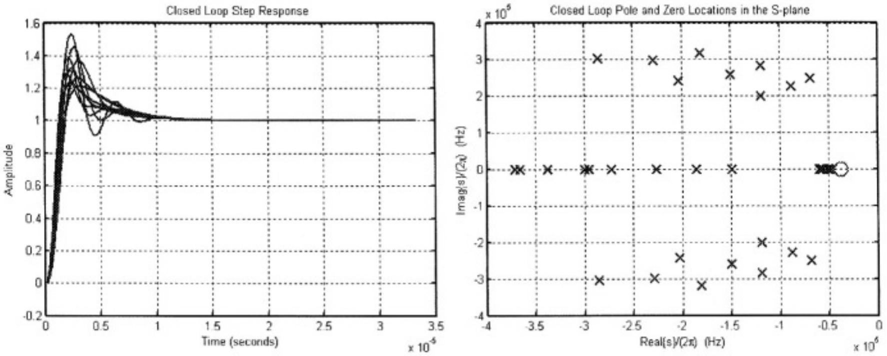


Fig.15. Resulting variations in the step response and pole/zero locations.

3. Synthesizer Simulation Techniques

We now turn our attention to the task of simulating fractional-N frequency synthesizers. Fast behavioral simulation of these systems is a challenging problem for a variety of reasons. First, the high output frequency of the synthesizer (often in the GHz range) imposes a high simulation sample frequency for traditional simulators. Unfortunately, the overall PLL dynamics have a bandwidth that is typically three to four orders of magnitude lower in frequency than the output frequency (often 100 kHz to 1 MHz bandwidth compared to a 1 to 10 GHz output frequency). Thus, traditional simulators take a long time to compute the dynamic response of the system since many simulation samples are required. This is the classical problem that is encountered with the simulation of PLL circuits. Second, for noise simulation, the fractional-N synthesizer adds the additional constraint that its behavior is non-periodic in steady-state due to the dithering action of the divide value, which prevents the use of fast methods developed for periodic steady-state conditions [10] as used with simulators such as SpectreRF.

Two techniques are presented here that allow fast and accurate simulation of both dynamic and noise performance of fractional-N synthesizers at a detailed behavioral level. The first provides accurate representation of the continuous-time (CT) PFD output with a discrete-time (DT) sequence using an area conservation principle. The second allows a dramatic reduction of the simulation sample frequency, and therefore a longer sample period, by including the divider implementation in the VCO simulation module. Both of these methods allow a uniform time sample period to be used, and also allow non-iterative computation of the sample values of the various signals within the system. The uniform time sample period allows the results of the simulator to be readily examined in the frequency domain without re-sampling, and the non-iterative computation allows the technique to be easily used in mainstream simulators such as Verilog, VHDL, Matlab, and custom C/C++ programs.

3.1. Proposed Area Conservation Method

Accurate simulation of fractional-N synthesizers is difficult due to the fact that the continuous-time (CT) signals of the synthesizer must be simulated in discrete-time (DT). The conversion process from CT to DT leads to the introduction of numerical noise for classical constant-time step simulation techniques. The key signal for introducing such noise is the PFD output. The standard approach of converting the CT PFD output to a DT sequence is to apply the sampling operation shown in Figure 16 [11]. Unfortunately, this approach effectively quantizes the location of the PFD edges according to the simulation sample period, $\mathbf{T_s}$. A reasonable assessment of the dynamic performance of the PLL can be achieved if $\mathbf{T_s}$ is made sufficiently small. However, the resulting quantization noise overpowers the true noise characteristics of the signals, and prevents proper noise analysis of the overall PLL.

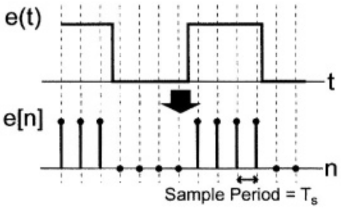


Fig.16. Classical uniform time sample period.

We suggest an alternative method for sampling the PFD signal that achieves accurate discretization of its output. To explain the technique, consider the illustration of this signal shown in Figure 17. Here we view the PFD output as a series of rectangular pulses with height of one or zero and a width and time offset that varies according to the location of PFD edges. For rectangular pulses not associated with edges, the width corresponds to the sample period of the simulation, $\mathbf{T_s}$. For rectangular pulses at edge boundaries, the width of the pulse varies between 0 and $\mathbf{T_s}$ as shown in the figure. In either case, these pulses look like impulses to the loop filter so that, from an intuitive standpoint, their influence can be characterized by two parameters – their area and time offset. Therefore, in line with this intuition, let us represent the corresponding DT PFD signal, e[n], as samples that have amplitude proportional to the *area* of the respective rectangular pulse in that time sample interval. The area of each pulse corresponds to its associated timing parameter ε shown in the figure --- the method of calculating ε for each pulse will be discussed in the next section. A mathematical justification for this technique is given in [5],

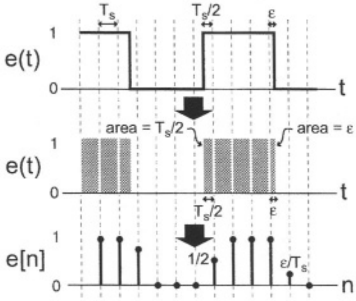


Fig.17. Details of PFD discretization technique.

Given the above PFD output discretization, we can then represent the CT loop filter dynamics with a discrete-time filter using either the impulse invariance or bilinear transform method [12]. This approach allows non-iterative computation of the loop filter dynamics. Figure 18 illustrates an example DT loop filter impulse response, along with a corresponding DT PFD signal. The charge pump is ignored in this analysis for simplicity; its effect can be included by simply scaling the PFD output by the value of the charge pump current. In the example, we see that the DT version of the loop filter simply consists of a DT filter whose impulse response corresponds to samples of the CT impulse response of the loop filter.

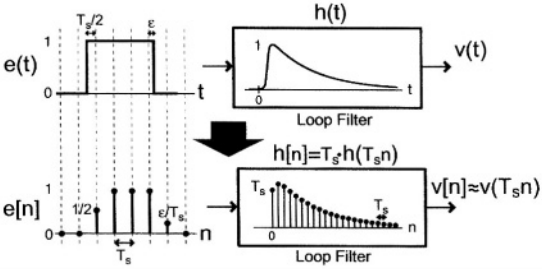


Fig.18. Discrete-time PFD output and associated loop filter response.

Now that it has been established that the PFD output signal can be accurately represented as a DT sequence using a principle of area conservation, let us examine the practical issue of implementing this technique in simulation code. Of particular concern is the ability to encompass a wide variety of PFD topologies. As revealed by Figure 19, computation of e[n] for a given PFD topology requires that transition information be passed along and processed by primitive elements such as registers and logic gates.

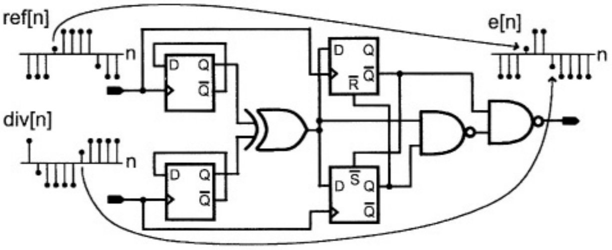


Fig.19. XOR-based PFD.

The transfer of transition information through primitives turns out to be straightforward, as illustrated in Figure 20 for a register and a representative logic gate, namely the 'and' gate. In the case of the register, the relevant timing information is contained in the clock signal. Specifically, whenever there is a transition at the output of the register, the location of that transition in time is set by the location of the rising (or falling) edge of the clock. As shown in the figure, this information is transferred to the register output by simply passing on the clk transition value when the output transitions in the same direction, and passing on the complement of the clk transition value when the output transitions in the opposite direction. In the case of the 'and' gate, either input can cause the output to transition. As gleaned from the figure, it is straightforward to determine which input is causing the transition, and appropriately pass its edge location value to the output of the 'and' gate. Similar arguments can be made for more complicated registers that include set and reset functions, and other primitives such as 'or' and 'xor' gates.

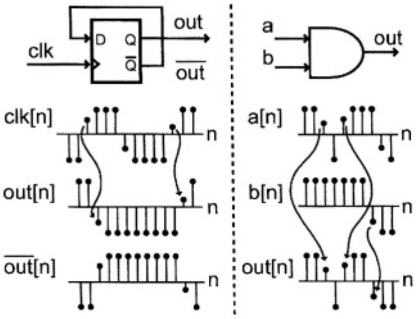


Fig.20. Example of transition value processing for a register and logic gate.

3.2 Fast Computation by Combining VCO and Divider Modules

To obtain fast computation, we propose a technique that allows us to set the simulation sample period, T_s, according to the reference frequency rather than the much higher VCO frequency. By doing so we can typically achieve more than two orders of magnitude speedup in simulation time of the PLL since the VCO frequency is typically more than two orders of magnitude higher than the reference frequency. This goal is achieved by combining the VCO and divider into one computation block.

To understand how we can combine the VCO and divider blocks, let us first consider the fact that the VCO phase can be modeling as an integrator fed by the VCO input [11]. We can, in turn, relate the rising edges of the divider output to the VCO phase by noting that such edges occur every time the VCO phase increments by $2\pi N[m]$ radians, where N[m] corresponds to the instantaneous divide value. Therefore, as illustrated on the left side of Figure 21, the VCO phase completely specifies the location of the divider edges. As such, we can determine the value of ε_k at the transition points of the divider output based entirely on computed VCO phase, as shown on the right side of Figure 21 for first-order interpolation [13]. By using VCO phase rather than the VCO output to compute the divider output, it suffices to choose a sample rate for the VCO phase computation according to the divider frequency, which equals the reference frequency, rather than the much higher VCO frequency. An empirically-based rule of thumb to apply when using this approach is to set the simulator sample rate (which corresponds the VCO phase computation sample rate) to be 6 to 20 times higher than the reference frequency.

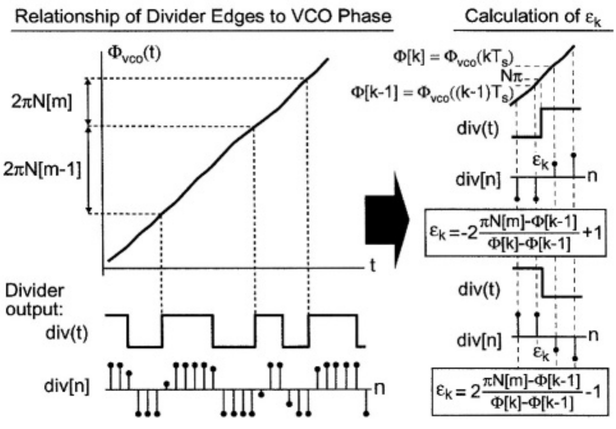

Relationship of Divider Edges to VCO Phase
Calculation of ε_k
$\Phi_{vco}(t)$
$2\pi N[m]$
$2\pi N[m-1]$
t
Divider output: div(t)
div[n]
n
$\Phi[k] = \Phi_{vco}(kT_s)$
$N\pi$
$\Phi[k-1] = \Phi_{vco}((k-1)T_s)$
div(t)
div[n]
ε_k
n
$\varepsilon_k = -2\frac{\pi N[m]-\Phi[k-1]}{\Phi[k]-\Phi[k-1]}+1$
div(t)
div[n]
ε_k
n
$\varepsilon_k = 2\frac{\pi N[m]-\Phi[k-1]}{\Phi[k]-\Phi[k-1]}-1$

Fig.21. Calculation of divider transition values from VCO phase.

3.3. Results

Using the presented techniques, the results of simulating the dynamic behavior and noise performance of a prototype synthesizer described in [14] are now discussed. The simulations were performed by the CppSim simulator package as described earlier in this chapter. To verify its accuracy, simulated noise will be compared to measured plots.

Figure 1 provides a block diagram of the prototype system; the reader is referred to [14] for more details. Relevant characteristics of the prototype include a reference frequency of 20 MHz, a VCO with carrier frequency of 1.84 GHz and K_v = 30 MHz/V, a second order $\Sigma-\Delta$ modulator, a charge pump that outputs +1.5 and -1.5 micro-Amps, a nominal divide value of 92.3, a PFD topology as shown in Figure 19 [8], and a lead/lag filter that achieves an overall closed loop PLL bandwidth of 84 kHz. Both dynamic and noise simulations will include the noise sources depicted in Figure 22, with VCO noise being input referred as a white noise source as described in [14]. Parameters associated with the noise sources that are shown in the figure were computed from Hspice simulations and VCO measurements.

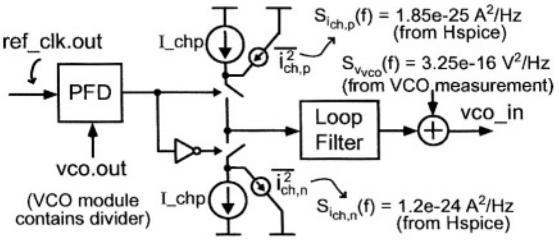

ref_clk.out
PFD
vco.out
(VCO module contains divider)
I_chp
I_chp
$\overline{i^2_{ch,p}}$
$\overline{i^2_{ch,n}}$
$S_{i_{ch,p}}(f) = 1.85e\text{-}25\ A^2/Hz$ (from Hspice)
$S_{v_{vco}}(f) = 3.25e\text{-}16\ V^2/Hz$ (from VCO measurement)
Loop Filter
vco_in
$S_{i_{ch,n}}(f) = 1.2e\text{-}24\ A^2/Hz$ (from Hspice)

Fig.22. Model of charge pump and VCO noise.

The simulation sample frequency was chosen as $1/T_s$ = 400 MHz, which is a factor of 20 higher than the reference frequency. The CT loop filter was converted to DT using the bilinear transform [12]. All simulations were run on a 650 MHz Pentium III laptop computer.

Beginning with dynamic behavior, Figure 23 shows the simulated VCO output frequency (constructed from the simulated VCO input) in response to variations at the input of the Σ–Δ modulator that include step and ramp functions. The step size is chosen to be large enough to knock the synthesizer out of frequency lock --- the corresponding oscillations in the VCO output frequency are a result of cycle slipping before the VCO becomes frequency locked again. The subsequent ramp in divide value illustrates the high resolution of the synthesizer as its output frequency is varied over a 40 MHz range. For this simulation, 260 thousand time steps were computed in less than 5 seconds.

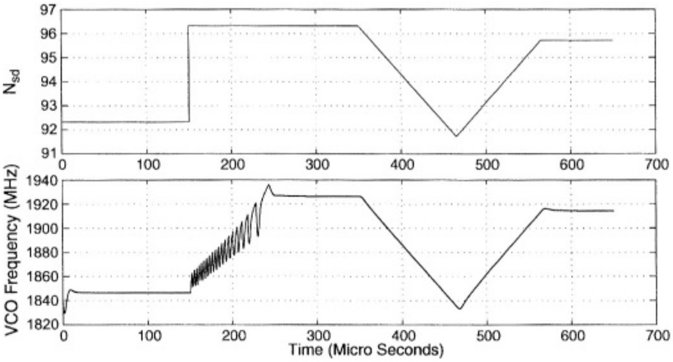


Fig.23. Simulated synthesizer dynamics.

A noise simulation of the prototype (constructed from the simulated VCO input) is shown in Figure 24 with the input to the **Σ–Δ** modulator being held constant. The plot shows the simulated output noise spectral density with the simulation sample frequency, **$1/T_s$**, set to 20 times the reference frequency. We see that the simulated phase noise agrees quite well with measured results. The larger discrepancy at frequencies close to 100 kHz is probably due to non-ideal characteristics of the charge pump, such as duty cycle offset and transient dynamics, not being modeled. Such effects could be included within the given framework, but it is useful to observe that, despite ignoring such effects, the simulation results are still quite accurate for this prototype. For this simulation, 5 million time steps were computed in 80 seconds.

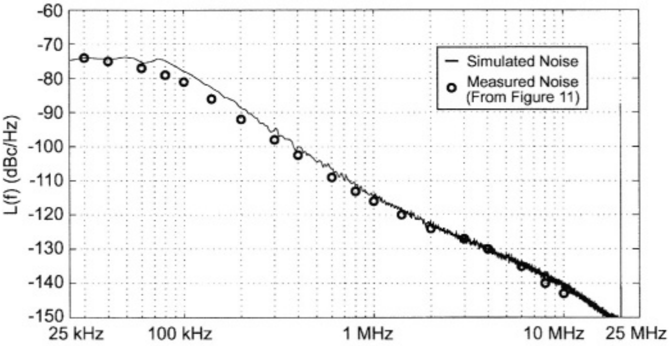


Fig.24. Simulated versus measured synthesizer phase noise.

4. Conclusions

In this paper we introduced design and simulation techniques for fractional-N frequency synthesizers. The design technique leverages a program called the PLL Design Assistant that allows fast and straightforward PLL design at the transfer function level. The simulation techniques are implemented in a program called CppSim, and leverage an area conservation approach and combining of the VCO and divider modules to achieve accurate and fast synthesizer simulation. Using these tools and techniques, designers can now explore both the dynamic and noise behavior of fractional-N synthesizer topologies at the behavioral level in a fast and efficient manner.

References

[1] T.A. Riley, M.A. Copeland, and T.A. Kwasniewski, "Delta-Sigma Modulation in Fractional-N Frequency Synthesis", Journal of Solid-State Circuits (JSSC), vol 28, no 5, pp 553-559, May 1993.

[2] M.H. Perrott, T. Tewksbury, and C. Sodini, "A 27 mW CMOS Fractional-N Synthesizer using Digital Compensation for 2.5 Mb/s GFSK Modulation", JSSC, vol 32, no 12, pp 2048-2060, Dec 1997.

[3] M.H. Perrott, M.D. Trott, C.G. Sodini, "A Modeling Approach for Sigma-Delta Fractional-N Frequency Synthesizers Allowing Straightforward Noise Analysis", JSSC, vol 38, no 8, pp 1028-1038, Aug 2002.

[4] Thomas H. Lee, "The Design of CMOS Radio-Frequency Integrated Circuits", Cambridge University Press, 1998.

[5] M.H. Perrott, "Fast and Accurate Behavioral Simulation of Fractional-N Synthesizers and other PLL/DLL Circuits", Design Automation Conference (DAC), 2002, pp 498-503.

[6] S. Willingham, M.H. Perrott, B. Setterberg, A. Grzegorek, W. McFarland, "An Integrated 2.5 GHz Sigma-Delta Frequency Synthesizer with 5 microseconds Settling and 2 Mb/s Closed Loop Modulation", International Solid-State Circuits Conference (ISSCC), 2000, pp 200-201.

[7] C.Y. Lau, M.H. Perrott, "Fractional-N Frequency Synthesizer Design at the Transfer Function Level Based on a Direct Closed Loop Realization Algorithm", DAC, 2003.

[8] A. Hill, A. Surber, "The PLL Dead Zone and How to Avoid It", RF Design, pages 131-134, Mar 1992.

[9] S. Norsworthy, R. Schreier, G. Temes, "Delta-Sigma Data Converters: Theory, Design, and Simulation", IEEE Press, 1997.

[10] K. Kundert, J. White, and A. Sangiovanni-Vincentelli, "Steady-State Methods for Simulating Analog and Microwave Circuits", Kluwer, Boston, 1990.

[11] D. Johns and K. Martin, "Analog Integrated Circuit Design", Wiley, 1997.

[12] A.V. Oppenheim and R.W. Schafer, "Discrete Time Signal Processing", Prentice Hall, NJ, 1999.

[13] A. Demir, E. Liu, A.L. Sangiovanni-Vincentelli, and I. Vassiliou, "Behavioral Simulation Techniques for Phase/Delay-Locked Systems", Custom Integrated Circuits Conference (CICC), pages 453-456, 1994.

[14] M.H. Perrott, "Techniques for High Data Rate Modulation and Low

Power Operation of Fractional-N Frequency Synthesizers with Noise Shaping”, PhD Thesis, MIT, 1997.

This page intentionally left blank

Monolithic CMOS Fractional-N Frequency Synthesizer Design for High Spectral Purity

Bram De Muer and Michiel Steyaert
Katholieke Universiteit Leuven, ESAT-MICAS
Kasteelpark Arenberg 10, 3001 Heverlee, Belgium
demuer@esat.kuleuven.ac.be

Abstract

Does fractional-N synthesis offer the way out for monolithic CMOS integration of high-quality transceivers? That is the question raised and answered in this document based on the most critical criteria influenced by extending the integer phase-locked loop (PLL) with fractional capabilities: phase noise and spurious tones, i.e. the spectral purity, integratability and agility. Linear system theory is applied to uncover the fundamental bandwidth limitations imposed by the $\Delta\Sigma$ noise in typical PLLs. Practice however proves the linear approach inaccurate. Therefore, a non-linear analysis method is developed, that swiftly predicts the effects of PLL non-linearities on the spectral purity. Serious in-band noise leakage and re-emerging spurious tones can be observed and are in close correspondence with experimental results. Both methods are applied to compare MASH and single-loop $\Delta\Sigma$ modulators in fractional-N synthesis. Based on the analyses, practical circuit design guidelines are compiled and applied to design a monolithic $\Delta\Sigma$-controlled fractional-N PLL in 0.25μm CMOS that complies to the stringent DCS-1800 cellular specifications.

1. Introduction

The dusk of the 20th and the dawn of the 21st century is characterized by the unrivaled growth of the telecommunication industry. The main cause was the introduction of digital signal processing in wireless communications, driven by the development of high performance, low cost, CMOS technologies for VLSI. However the implementation of the RF analog front-end remains a bottleneck. This is reflected in the large effort put in monolithic CMOS integration of RF circuits both by academics and industry [1–3].

The goal of this work is the monolithic integration in standard CMOS technology of a frequency synthesizer to enable the full integration of a transceiver front-

A. van Roermund et al. (eds.), Analog Circuit Design, 51-75.
© 2003 *Kluwer Academic Publishers. Printed in the Netherlands.*

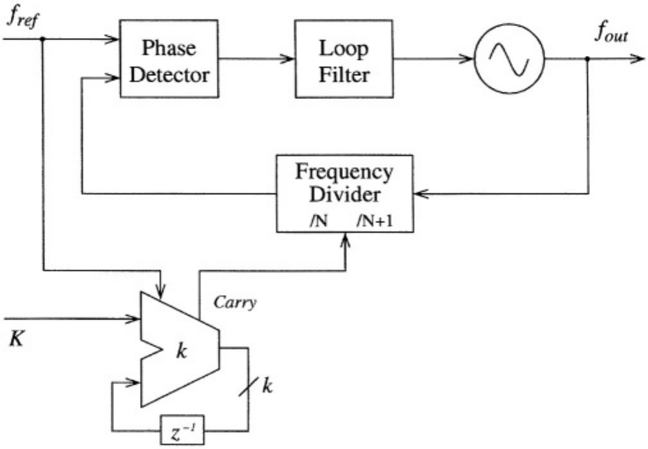


Figure 1: The Fractional-N Principle.

end in CMOS [1]. The $\Delta\Sigma$ fractional-N synthesizer topology presents an elegant solution [4]; It circumvents the severe speed-spectral purity-resolution trade-off of the classic PLL synthesizer by providing synthesis of fractional multiples of the reference frequency.

The basic idea behind fractional-N synthesis is division by *fractional* ratios, instead of only integer ratios [5]. This is accomplished by e.g. dividing 99.9% of the time by N and 0.1% by $N + 1$, such that the mean division is by $N.001$. To accomplish fractional division, the same frequency divider as in an integer-N frequency synthesizer is employed, but the division is controlled differently. In Fig. 1 the division modulus of the frequency divider is steered by the carry output of a simple digital accumulator of k-bit width. To realize a fractional division ratio $N + n$, with $n \in \mathbb{R}[0, 1]$, a digital input $K = n \cdot 2^k$ is applied to the accumulator. A carry output is produced every K cycles of the reference frequency f_{ref}, which is also the sampling frequency of the digital accumulator. This means that the frequency divider divides $2^k - K$ times by N and K times by $N + 1$, resulting in a division ratio N_{frac}, given by Eq. (1).

$$\begin{aligned} N_{frac} &= \frac{\left(2^k - K\right) \cdot N + K \cdot (N+1)}{2^k} \\ &= N + \frac{K}{2^k} = N + n \end{aligned} \tag{1}$$

Eq. (1) states that for a given reference frequency, it is possible to make the fre-

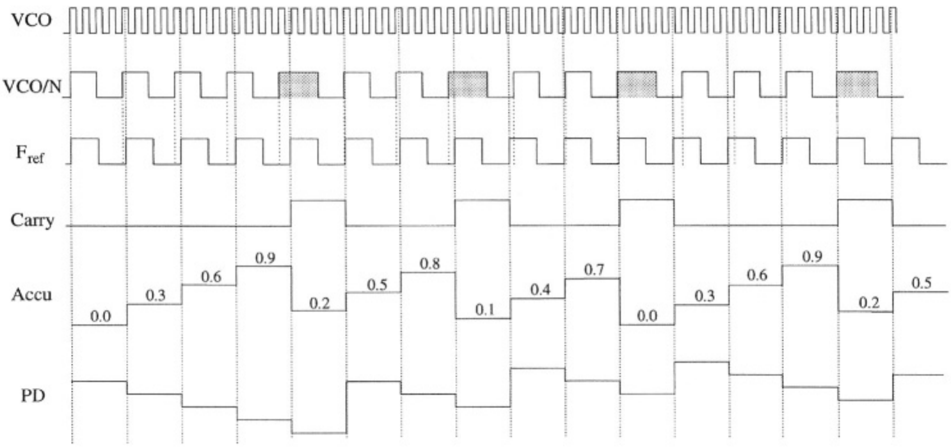


Figure 2: The chronograph of the VCO and frequency divider output, the overflow bit, the accumulator output and the phase detector output for $N = 4$ and $n = 0.3$.

quency resolution arbitrary fine, by choosing the width of the accumulator sufficiently large. For example, in a DCS-1800 telecommunication system the channel spacing of 200 kHz can be synthesized using a f_{ref} of 26 MHz, by realizing an accumulator width k of more than 7 bits. Each cycle of the reference clock, the accumulator accumulates the input K, until the accumulator output overflows. At overflow, the carry output is set high and 2^k is subtracted from the accumulator output. The resulting accumulator output is a sawtooth waveform with a frequency depending on n, as shown in Fig. 2 for $n = 0.3$. The carry output of the accumulator is used to control the division ratio of the divider. The carry output has a mean value of n, but the instantaneous value is merely a "prediction" of the mean value, since the output can only be 0 or 1. Therefore, the frequency divider is dividing only by integer ratios, N or $N + 1$, while the output of the PLL is a fractional multiple of f_{ref}. When the divider is dividing by N, the divider output leads the reference frequency, giving rise to an increasing phase difference between both phase detector inputs. The phase detector output current starts to accumulate. When the accumulator overflows, the divider modulus is changed from N to $N + 1$, meaning that one output period of the VCO is swallowed by the frequency divider (the shaded pulses in Fig. 2). This causes a phase shift at the input of the phase detector that eliminates the previously accumulated phase error. Fig. 2 shows the chronograph of the open-loop output of the phase detector versus the VCO output and f_{ref} for $N = 4$ and $n = 0.3$.

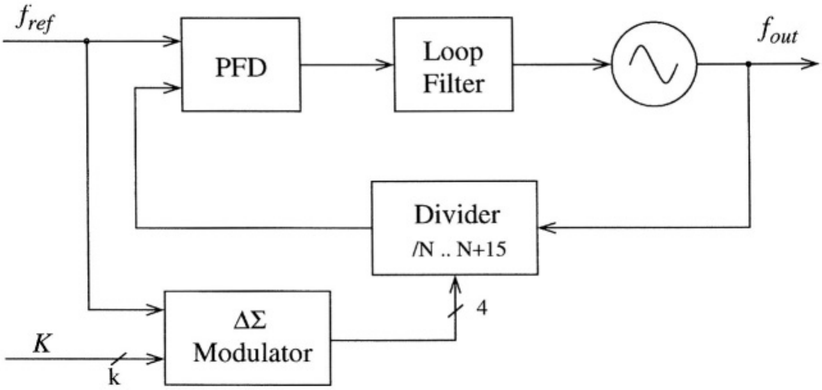


Figure 3: The principle of $\Delta\Sigma$ *fractional-N synthesis.*

The accumulator, as shown in Fig. 1, acts as a *phase* accumulator; The phase error in the PLL is in fact a time-integrated scaled version of the output of the digital accumulator. The register is also called the *phase register.* Note that the fractional-N PLL is never really locked, since the phase error is never zero over more than one period of f_{ref}. The pulse swallowing action, performed by the frequency divider/accumulator circuit, is a periodic action, i.e. every $2^k - K$ cycles of f_{ref} a VCO pulse is swallowed. Therefore, the sawtooth-shaped phase error is also periodic with a period depending on K. The resulting AC component, superimposed on the wanted DC output of the phase detector, modulates the VCO frequency and large spurious components appear in the PLL output spectrum at multiples of $n \times f_{ref}$, even after filtering by the loop filter.

Since the accumulator is in fact nothing more than a first-order $\Delta\Sigma$ modulator, $\Delta\Sigma$ theory can be applied to the fractional-N synthesizer (see Fig. 3). The $\Delta\Sigma$ modulator output controls the instantaneous division modulus of the prescaler, such that the mean division modulus is $N_{frac} = N + K/2^k$, with k the number of bits of the $\Delta\Sigma$ modulator and K the input word. The corresponding phase changes at the prescaler output are quantized, leading to possible spurious tones and quantization noise. By selecting higher order $\Delta\Sigma$ modulators, the spurious energy is whitened and shaped to high frequency noise, which can be removed by the low-pass loop filter. As a result, for a given frequency resolution, an arbitrary high reference frequency, f_{ref} can be chosen, ideally without degrading the spectral purity of the synthesizer. Therefore, the $\Delta\Sigma$ fractional-N frequency synthesizer offers the most integrated solution and the best path to full CMOS integration:

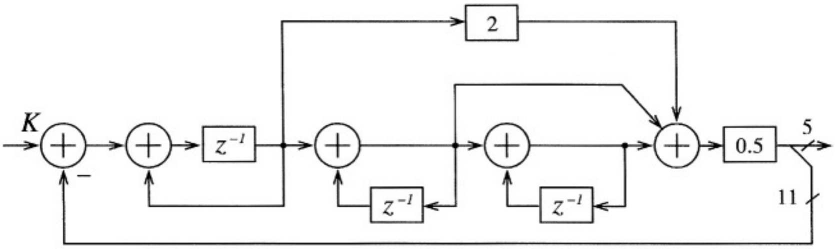


Figure 4: The 3rd-order multi-bit, single-loop $\Delta\Sigma$ modulator. From the 5 output bits only 4 are used for stability reasons.

- Due to the higher reference frequency, fractional-N synthesis offers a potentially higher bandwidth, enabling the full integration of the loop filter, without blowing up the IC area.
- In a $\Delta\Sigma$ fractional-N synthesizer frequency resolution is traded against digital complexity, which is the natural biotope of CMOS technology.
- The higher reference frequency lowers the necessary division factor, such that the noise in the PLL band is less amplified; As a result the in-band noise, and thus the rms phase error, can be lowered.
- The broader bandwidth enables agile frequency switching, necessary for high data rate communication systems.
- The $\Delta\Sigma$ fractional-N PLL requires only one PLL and a digital $\Delta\Sigma$ modulator, permitting power and area savings.

2. $\Delta\Sigma$ Fractional-N Synthesis: Linear Analysis

To model the impact of $\Delta\Sigma$-control on the spectral purity of the synthesizer, a linear-time-invariant (LTI) PLL model is employed, with the $\Delta\Sigma$ quantization noise as an additive noise source at the prescaler output. The prescaler with $\Delta\Sigma$-control can be looked upon as a digital-to-phase (D/P) converter. Every reference cycle, the prescaler subtracts $n \cdot 2\pi$ rad from its input signal, with $n = 0..(2^b - 1)$ determined by the $\Delta\Sigma$ modulator output. As guinea pigs for the analysis, a 3rd-order MASH modulator and the 3rd-order, single-loop, multi-bit modulator of Fig. 4 are chosen, since they represent the far ends of the $\Delta\Sigma$ modulator spectrum. The single-loop modulator contains only 1 pole at the origin of the z-plane and two low-Q Butterworth poles at $0.167 \times f_{ref}$, with a passband gain of 3.2. The resulting quantization noise on the division modulus, and thus output phase is approximated by uniformly distributed white noise [6]. The quantization noise power is $\Delta^2/12$ with $\Delta = \Delta N/(2^b - 1) = 1$ for both modulators with ΔN the modulus range and b the number of effective $\Delta\Sigma$ output bits.

Out-of-Band Phase Noise

The out-of-band phase noise contribution of the $\Delta\Sigma$ modulator at the output of the synthesizer is found in Eq. (2) [7, 8], with $H_{qn}(z)$ the $\Delta\Sigma$ noise transfer function. The noise transfer function is integrated since the prescaler acts on frequency, while the state variable is phase.

$$S_{\Delta\Sigma,o}(f) = \frac{\pi^2}{3 \cdot f_{ref}} \cdot \frac{|H_{qn}(z)|^2}{|1 - z^{-1}|^2} \cdot |T(f)|^2 \tag{2}$$

In the equation, $T(f)$ is the closed-loop transfer function of the PLL. Since the aim is to achieve high spectral purity, a 4th-order, type-II PLL characteristic is chosen (see Eq. (3)), The open-loop gain can be approximated by $\beta^2\,(f_c/f)^3$ in the frequency range $f_c \ll f < f_{ref}$, with β the relative position of the two HF poles relative to the cross-overfrequency, f_c, to ensure enough phase margin. The transfer function of the PLL becomes then:

$$|T(f)|^2 = \frac{\beta^4 f_c^6}{f^6 + \beta^4 f_c^6 + 2\beta^2 f_c^3} \cong \beta^4 \frac{f_c^6}{f^6} \tag{3}$$

The transfer function of Eq. (3), is in fact an approximation of a 3rd-order Butterworth filter with one pole at f_c and two poles at βf_c. Note that the order of the chosen $\Delta\Sigma$ modulators must be equal or lower than that of the loop filter (three in this case) to ensure an equal or steeper noise roll-off ($\leq$-20dB/dec) at the frequencies of interest.

Since the main advantage of $\Delta\Sigma$ fractional-N synthesizers is the decoupling of the reference frequency f_{ref} and the PLL bandwidth f_c, the influence of the $\Delta\Sigma$ noise on the bandwidth requirement is examined. Substituting the $|T(f)|^2$ in Eq. (2) leads to the formula for the maximum bandwidth of the PLL for an n-th-order MASH modulator:

$$f_{c,max} = \left[S_{\Delta\Sigma,o}(f) \cdot \frac{12}{(2\pi)^{2n}} \cdot f_{ref}{}^{2n-1} \cdot \frac{f^{8-2n}}{\beta^4} \right]^{1/6} \tag{4}$$

To comply with the most stringent DCS-1800 phase noise spec, i.e. -133 dBc/Hz at 3 MHz offset [9], the target $\Delta\Sigma$ phase noise is $S_{\Delta\Sigma,o}(3\text{ MHz}) = -136$ dBc/Hz. In Fig. 5, the maximum PLL bandwidth, $f_{c,max}$ is plotted versus the reference frequency for different MASH modulator orders.

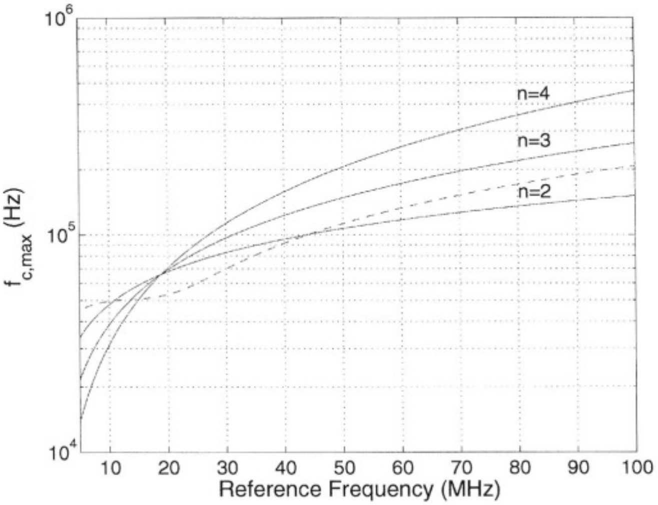


Figure 5: The maximum PLL bandwidth $f_{c,max}$ versus the reference frequency and different $\Delta\Sigma$ modulator orders, for a type-II, 4-th order PLL. The dashed curve is for the 3rd-order, single-loop modulator. The targeted phase noise specification is -136 dBc/Hz at 3 MHz for DCS-1800.

The dashed line is the maximum bandwidth for a single-loop, multi-bit $\Delta\Sigma$ modulator Fig. 4. Due to the Butterworth poles, the noise power is approximately a factor 4 higher in the frequency band of interest than for a MASH modulator. For a reference frequency of 26 MHz, not much is gained from increasing the modulator order. For a high bandwidth, the reference frequency and/or the modulator order should be increased leading to an increased power consumption and circuit complexity. The maximum bandwidth is 87 kHz for the 3rd-order MASH modulator and only 62 kHz for the single-loop, multi-bit modulator. Due to the stringent noise specifications of high-quality wireless standards, the possible bandwidth of a fractional-N synthesizer is severely limited.

rms Phase Error

Apart from the out-of-band phase noise constraint, the integrated in-band phase noise, determining the rms phase error $\Delta\Phi_{rms}$ of the PLL is of importance. To be sure the $\Delta\Sigma$ does not corrupt the rms phase error, the dynamic range of the modulator must be higher than the dynamic range of the PLL [10]. The integrated in-band frequency noise Δf_n is given by $\Delta f_n^2 \approx 2/3 \cdot A_n \cdot f_{nbw}^3$ with f_{nbw} the

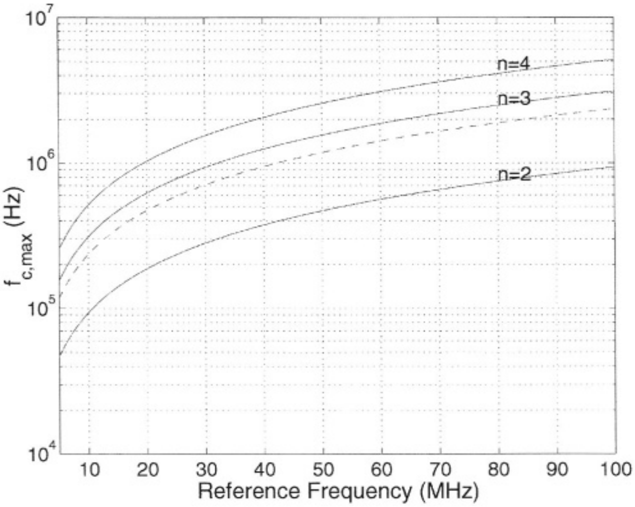


Figure 6: The maximum PLL bandwidth $f_{c,max}$ versus the reference frequency and different $\Delta\Sigma$ modulator orders for $\Delta\Phi_{rms} < 1.5^{\circ}$. The dashed curve is for the 3rd-order, single-loop modulator.

noise bandwidth of the PLL and $10 \log A_n$ the in-band phase noise in dBc/Hz. The noise bandwidth of a 4th-order, type-II PLL is $f_{nbw} \approx 1.16 f_c$. From this, the dynamic range of the frequency synthesizer can be derived:

$$DR_{PLL} = \left(\frac{\Delta N \cdot f_{ref}}{\Delta f_n}\right)^2 = 8 \cdot \frac{\Delta N^2}{\Delta\Phi_{rms}^2} \cdot \left(\frac{f_{ref}}{2 f_c}\right)^2 \tag{5}$$

The dynamic range of an n-th-order MASH modulator is [6]:

$$DR_{\Delta\Sigma} = \frac{3}{2} \cdot (2^b - 1)^2 \cdot \frac{2n+1}{\pi^{2n}} \cdot \left(\frac{f_{ref}}{2 f_c}\right)^{2n+1} \tag{6}$$

Calculating the inequality gives the maximum bandwidth f_c of the PLL in Eq. (7).

$$f_c < \left[\frac{3}{8} \cdot \frac{2n+1}{(2\pi)^{2n}} \cdot \Delta\Phi_{rms}^2\right]^{\frac{1}{2n-1}} \cdot f_{ref} \tag{7}$$

The maximum PLL bandwidth $f_{c,max}$ is plotted versus the reference frequency of the PLL for different MASH modulator orders in Fig. 6 and for a 1.5° rms phase error (to ensure at least an overall rms phase error of 2°). For the single-loop, multi-bit $\Delta\Sigma$ modulators (dashed curve), the actual maximum bandwidth

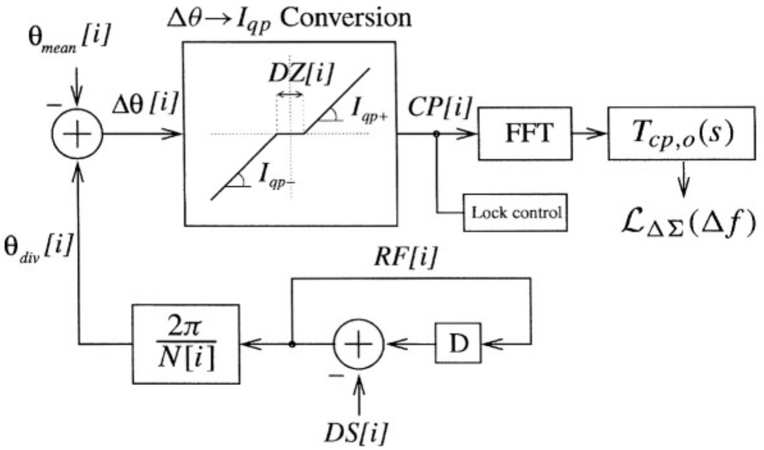


Figure 7: The schematic representation of the non-linear, transient analysis method with all corresponding simulation parameters and variables.

is again 25% smaller than in Eq. (7). In the case of a 3rd-order modulator and $f_{ref} = 26\text{MHz}$, the maximum bandwidth is 810 kHz and 614 kHz, respectively. According to linear theory, the constraint posed on the $\Delta\Sigma$ modulator noise due to in-band noise contributions is much less severe than the constraint due to the out-of-band phase noise at 3 MHz [1].

3. $\Delta\Sigma$ Fractional-N Synthesis: Non-Linear Analysis

The linear analysis suggested that applying $\Delta\Sigma$ control to the prescaler would not cause any problems whatsoever. Practice however proves this inaccurate. Therefore, an analysis method is developed that provides a fast, interactive means of examining the influence of non-linearities and mismatch in different PLL building blocks by performing a transient simulation of the $\Delta\Sigma$ fractional-N synthesizer. The strong point of the analysis method is its capability to sweep simulations sufficiently fast over different degrees of non-linearities and operating points, while performing sufficiently long transient simulations to get accurate fast-Fourier transforms (FFT) of the phase variable. This is accomplished by analyzing the fractional operation of the PLL in discrete time and in open-loop. The building blocks are represented by high level models with parameters to model any non-linear behavior or mismatch in critical transistors as shown in Fig. 7.

---

[1] A more detailed version of the above analyses can be found in [8].

The jitter on the reference clock and thus on the $\Delta\Sigma$-generated modulus control pulses can also be incorporated in the analysis. The simulations are performed in Matlab [11].

To find the phase error, generated by the $\Delta\Sigma$ modulation of the division modulus, the variation of the number RF pulses, $RF[k]$, at the output of the divider is monitored. Every reference cycle, the number of RF pulses at the divider output is determined by the number of pulses swallowed by the $\Delta\Sigma$ control, $DS[k]$:

$$RF[i] = \sum_{k=2}^{i} RF[k-1] - DS[k] \tag{8}$$

The resulting quantized phase changes are compared with the phase that would be expected when the loop would be in lock, i.e. the phase corresponding to the fractional part of the division modulus $K/2^k$. The result is the instantaneous accumulated phase error $\Delta\theta[i]$ (Eq. (9).

$$\Delta\theta[i] = \sum_{k=2}^{i} \frac{2\pi}{N[k]}\left(RF[k] - (k-1).\frac{K}{2^k}\right) \tag{9}$$

The phase error is converted to current pulses, $CP[i]$, in the charge pump. The $\Delta\theta \rightarrow I_{qp}$ (phase error-to-charge pump current) conversion is modeled to contain any phase/frequency detector (PFD) non-linearity (see Fig. 7). Mismatch in the up and down current sources, resulting in gain mismatch for positive and negative phase errors is modeled by $I_{qp\pm}$. The occurrence of a dead zone is modeled by $DZ[k]$.

$$CP[i] = \sum_{k=2}^{i} \frac{I_{qp\pm}}{2\pi}\left(\Delta\theta[k] \pm DZ[k]\right) \tag{10}$$

The current noise spectrum is modeled as a phase noise source which is subjected to its corresponding closed loop transfer function. The non-linear conversion from voltage to frequency/phase in the VCO is modeled by the variation of the VCO gain, when changing the operating point of the PLL. The total analysis time is approximately 15 seconds on a Pentium4.

The analysis tool enables the evaluation and comparison of the effect of MASH and single-loop $\Delta\Sigma$ noise on the PLL. An analysis is performed with the following non-linearities; a 0.1% dead zone (0.72° around 0° phase error) and a gain mismatch of ±2%. The reference frequency is 26 MHz and the fractional division number is 67.92. The output frequency is 1.76592 GHz, i.e 2.08 MHz offset from an integer multiple of f_{ref}. In Fig. 8 (c) and (d), the time-domain phase error $\Delta\theta[i]$ is plotted for both modulators. Note that the $\Delta\Sigma$ fractional-N PLL

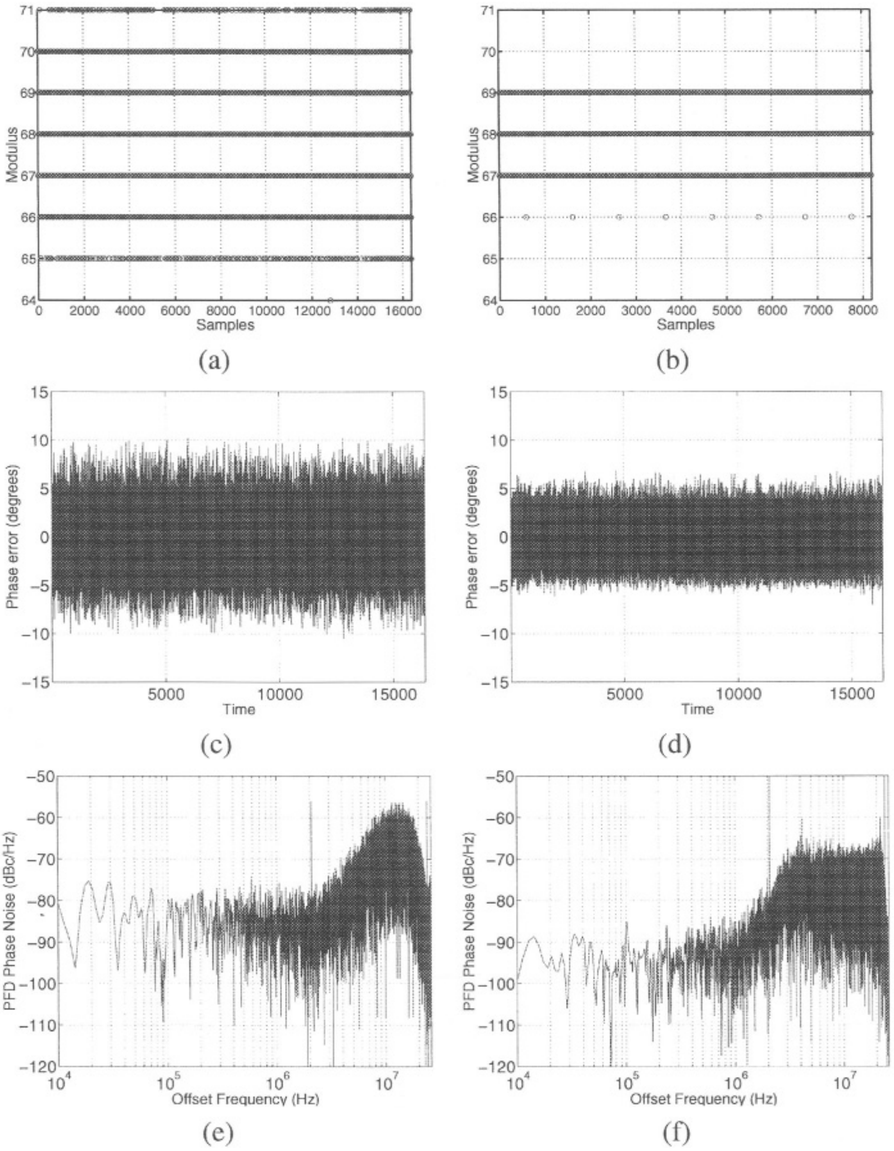


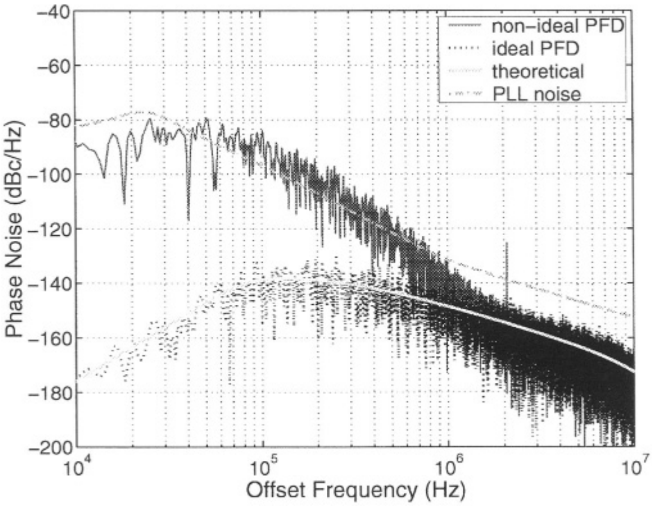
Figure 8: Simulation results: The division moduli of (a) the MASH and (b) the single-loop modulator. The phase error $\Delta\theta$ for (c) the MASH and (d) the single-loop modulator. The FFT of the current pulses $CP[i]$ for (e) the MASH and (f) the single-loop modulator.

frequency synthesizer can hardly be called a *phase-locked* loop, since the loop is never in lock! Fig. 8 (a) and (b) shows the time-domain modulus switching in the prescaler. The more intensive switching of the MASH modulator, which is reflected in the frequency domain by increased high frequency noise, is the cause of the larger instantaneous phase error. Due to the shaping of the HF noise in the single-loop $\Delta\Sigma$ modulator, i.e. less modulus switching, the instantaneous phase error is smaller. This has two important consequences; First, the on-time of the charge pumps is smaller for the single-loop modulator, making it less sensitive to noise coupling from the substrate and the power supply. Secondly, the noise leakage is reduced.

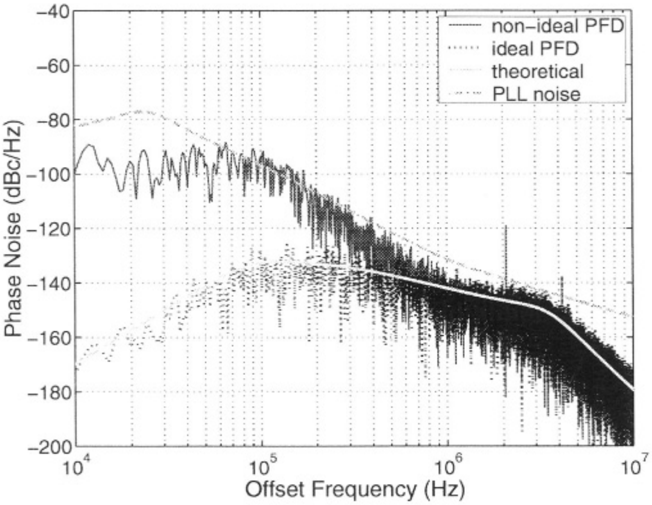
The FFTs of the charge pump current pulses $CP[i]$ are plotted in Fig. 8 (e) and (f). A noise floor appears in the output spectrum as well as spurious tones, although the $\Delta\Sigma$ output is perfectly randomized and dithered. Due to the non-linear mixing in the PFD-charge-pump, noise at $f_{ref}/2$ folds back to lower offset frequencies, similar to the effect of a non-linear DAC in a multi-bit $\Delta\Sigma$ ADC. Since the noise at $f_{ref}/2$ is much lower for the single-loop $\Delta\Sigma$ modulator, its noise leakage due to the non-linear mixing in the PFD is also lower. The difference in phase error between MASH and single-loop modulators is reflected in a 10 dB lower noise floor. In addition, spurious tones appear in the output spectrum at $j \times K/2^k \cdot f_{ref}$ with $j = 1,2,3,\ldots$

Fig. 9 shows the $\Delta\Sigma$ noise of both modulators as it appears at the PLL output for an ideal (dotted) and a non-linear $\Delta\theta \rightarrow I_{qp}$ conversion (solid). The results of the ideal case closely match the results of linear analysis (solid light grey). Due to non-linearity, the simulated output spectrum of the integer-N PLL [8, 12] (the dashed-dotted line) is seriously deteriorated by $\Delta\Sigma$ noise in the PLL noise bandwidth, increasing the $\Delta\Phi_{rms}$. Especially, the MASH converter is critical in terms of in-band noise due to the higher phase error, despite the inherently lower LF $\Delta\Sigma$ noise of the MASH modulator. Note that the simulations are performed without taking into account noise coupling through the substrate or power supply lines. As a consequence, the actual spurious performance of the $\Delta\Sigma$ factional-N PLL could be worse than simulated. The presented simulation results are for a division modulus 67.92, close to an integer multiple of f_{ref}. When analyzing division moduli in between integer multiples of f_{ref}, noise leakage is still observed, but the spurious tones are well below the phase noise.

The explanation for the re-emerging of spurious tones is that the modulator is unable to sufficiently decorrelate the successive output samples, especially for inputs close to an integer value. A second possible source of tones is the down-conversion of tones which are inherently present around $f_{ref}/2$ [6], by the non-linear mixing in the PFD. This effect can be worsened by substrate and power



(a)



(b)

Figure 9: Simulation results: The $\Delta\Sigma$ noise at the output of the PLL for (a) the MASH and (b) the single-loop modulator. The results are plotted for an ideal PFD (dotted), which closely corresponds to the linear analysis (solid light grey) and for a non-linear PFD (solid). They are compared to the simulated integer PLL phase noise (the dash-dotted line).

supply coupling with signals at $f_{ref}/2$.

An analysis sweep is performed, to study the influence of different degrees of non-linearity on the output phase noise of the synthesizer. In Fig. 10, the results of a sweep over different degrees of dead zone and gain mismatch on the phase noise are plotted. For the non-linearities, a log-scale is applied to show the linear relationship of the noise leakage and the non-linearities. Due to the high simulation speed, the analysis can be swept over different degrees of non-linearity and for different input values and the effects on phase noise as well as spurs can be monitored, leading to the following conclusions:

- The $\Delta\Sigma$ phase noise at 600 kHz is lower for a MASH modulator in the ideal case, but only a small non-linearity is enough to favor the single-loop modulator.
- The phase noise at 3 MHz is more critical for a single-loop modulator, due to the pole shift in the noise transfer function. The noise at 3 MHz is virtually not influenced by non-linearities.
- The in-band noise leakage for a single-loop $\Delta\Sigma$ modulator is always lower than for a MASH modulator. For the MASH modulator, noise leakage is highly critical in the presence of a dead zone.
- The influence of a very small dead zone is much more severe than the influence of gain mismatch.
- The on-time of the PFD is higher for the MASH than for the single-loop modulator, which increases the sensitivity of the synthesizer to digital noise coupling through the substrate and the power supply.
- The value of the fractional division input number has no direct influence on the noise leakage due to non-linearities.
- For fractional division by numbers in between integer division moduli, no spurious tones are present in the output spectrum.
- Spurious tones show up at the synthesizer output at the fractional frequencies, although dithering is applied. The spurious tones are higher when the synthesized frequency approaches an integer multiple of the reference frequency.

4. $\Delta\Sigma$ Fractional-N Synthesis: Example

As an example, a 4th-order, type-II PLL is integrated, including a 4-bit prescaler, a zero-dead-zone PFD, dual charge pumps and a 3-step equalizer together with an on-chip LC-tank VCO and a 35 kHz dual-path loop filter (see Fig. 11). The equalizer performs a 3-step piece-wise equalization of the loop gain. Only the

Loop Parameters		**Loop Passives**		**Performance (at 600 kHz)**	
ω_c	35 kHz	R_p	3.2 kΩ	$\mathcal{L}_{qp}$	-166.4 dBc/Hz
I_{qp}	2 μA	R_4	1.07 kΩ	$\mathcal{L}_{qp,B}$	-150.5 dBc/Hz
B	12	C_p	240 pF	$\mathcal{L}_{add}$	-133.0 dBc/Hz
f_{ref}	26 MHz	C_4	710 pF	$\mathcal{L}_{int}$	-137.0 dBc/Hz
N	67.92	C_z	450 pF	$\mathcal{L}_{R_p}$	-134.6 dBc/Hz
		C_{tot}	1.4 nF	$\mathcal{L}_{R_4}$	-132.5 dBc/Hz
				$\mathcal{L}_{tot,int}$	-124.3 dBc/Hz
				$\mathcal{L}_{MASH}$	-123.6 dBc/Hz
				$\mathcal{L}_{SL}$	-128.0 dBc/Hz
PM	55^o	L_{ind}	2.857 nH	$\mathcal{L}_{tot,MASH}$	**-120.9 dBc/Hz**
ζ	0.72	Q_L	9	$\mathcal{L}_{tot,SL}$	**-122.8 dBc/Hz**

Table 1: Summary of the loop properties and simulated performance of the $\Delta\Sigma$ fractional-N synthesizer at 1.76592 GHz.

loop filter and PFD-charge pump design is discussed[2] because of their crucial role in determining the spectral purity, the integratability and the agility. In Table 1, the most important loop properties are summarized. The fully integrated CMOS LC-VCO achieves a noise spec of less than -127.5 dBc/Hz at 600 kHz with a tuning range of over 28% [13]. To enable testing of different $\Delta\Sigma$ modulators, the $\Delta\Sigma$ modulator is not integrated and measurements are performed by applying a $\Delta\Sigma$ bit-stream to the prescaler inputs.

4.1. Monolithic Dual-Path Loop Filter Design

To achieve full integration, a dual-path filter topology has been implemented (Fig. 11). The addition of one active integration (C_z) and one passive low-pass filter path (C_p, R_p) with a scaling factor B realizes the low frequency zero for loop stability in a type-II PLL, without adding the actual capacitor [14]. The total number of capacitors is the same as in a classical 4th-order, type-II PLL, but for the same phase noise the total integrated capacitance is more than 5 times smaller [8].

A systematic approach is developed to determine the loop bandwidth and the charge pump current under the following constraints: $C_{tot} \leq 2\text{nF}$, $\mathcal{L}\{600\text{kHz}\} \leq$

---

[2] Detailed information on the design of the different building blocks can be found in [8].

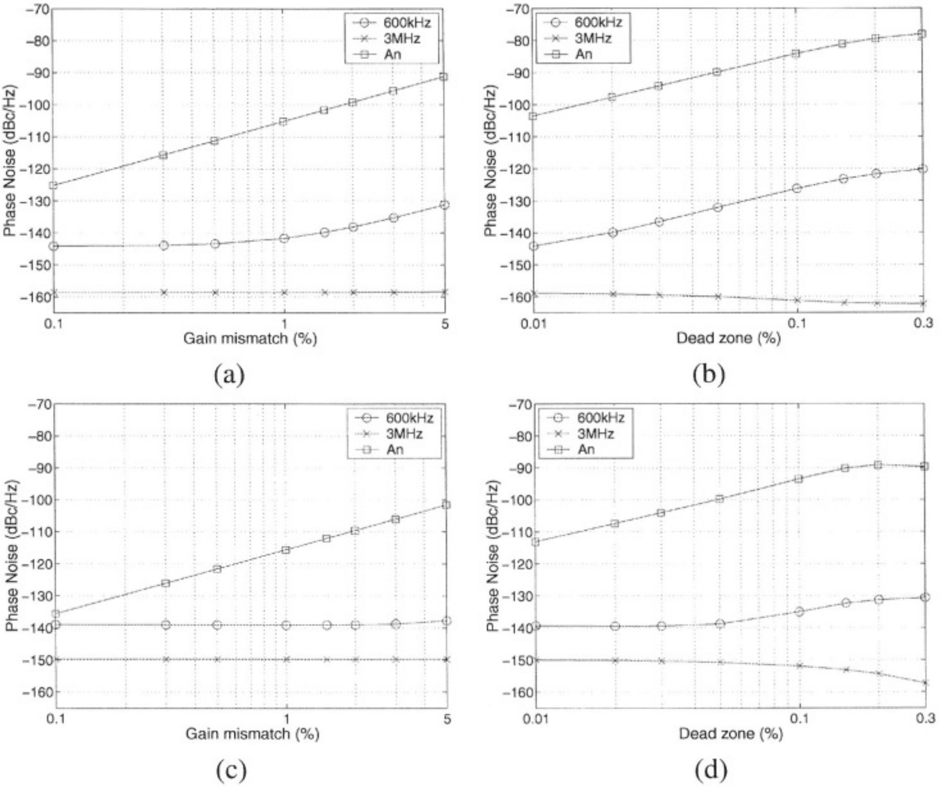
-123 dBc/Hz, $\Delta\Phi_{rms} \leq 2^{\circ}$ and $T_{settle} \leq 300\mu s$. In Fig. 12 (a) the corresponding design space is marked out for both ends of the DCS-1800 frequency band. At $N = 72$, the VCO gain is low and a large bandwidth can be chosen. However, at the low frequency side, the VCO gain is high (300 MHz/V) and the design space is significantly reduced (dashed). Therefore, the bandwidth needs to be quite low, i.e. 35 kHz (marker in Fig. 12 (a)) with 1.4 nF of integrated capacitance. Despite the rather low loop bandwidth for a fractional-N synthesizer, a settling time of less than 293 μs for a 104 MHz step is simulated with Saber [15] (Fig. 12 (b)). From this analysis, all pole and zero positions are determined, trading off low noise and high integratability [12].

4.2. PFD and Charge Pumps

To minimize noise leakage and spurious tones, the PFD and charge pump circuits are carefully optimized towards spurious suppression as such and towards a highly linear phase error detection (see Section 3.).

To minimize reference spurs, the integration in the first path of the loop filter is done actively to keep the charge pump output at a fixed level (see Fig. 11). The current switches arc implemented with PMOS and NMOS transistors to compensate charge injection. Furthermore, a timing control scheme (Fig. 11) delivers synchronized control signals to drive both the output current branch and the dummy current branch of the charge pumps, such that the current I always flows, preventing hard on/off switching of the current sources [8, 16].

To linearize the $\Delta\theta \rightarrow I_{qp}$ conversion, the phase detection is performed by a zero-dead-zone PFD [17], to prevent a hard non-linearity around 0° phase error. The on-time fraction of the charge pump due to the additional delay in the PFD is 7.5%. This value is a trade-off between dead-zone prevention and sensitivity to noise coupling. To further minimize digital noise coupling, the sampling in the PFD and the computational events in the $\Delta\Sigma$ modulator and prescaler are offset in phase. To make sure that the gains for positive and negative phase error detection are equal, the current source transistors are oversized to ensure sufficient matching. This also decreases the current source 1/f noise that affects the in-band noise. Additionally, the timing control provides synchronization between the two filter paths and the switches of the charge pumps themselves, thereby ensuring equal positive and negative phase error detection gain. Hspice simulations of $\Delta\theta \rightarrow I_{qp}$ circuits show no dead zone and a worst-case gain mismatch of ±0.5%.



(a) (b)

(c) (d)

Figure 10: Simulation results: The simulated output phase noise at 600 kHz, 3 MHz and in-band noise A_n, respectively, versus the gain mismatch in the $\Delta\theta \rightarrow I_{qp}$ conversion in ±% and the dead zone in % of the PFD phase range for the MASH $\Delta\Sigma$ modulator ((a) and (b)) and the single-loop $\Delta\Sigma$ modulator ((c) and (d)).

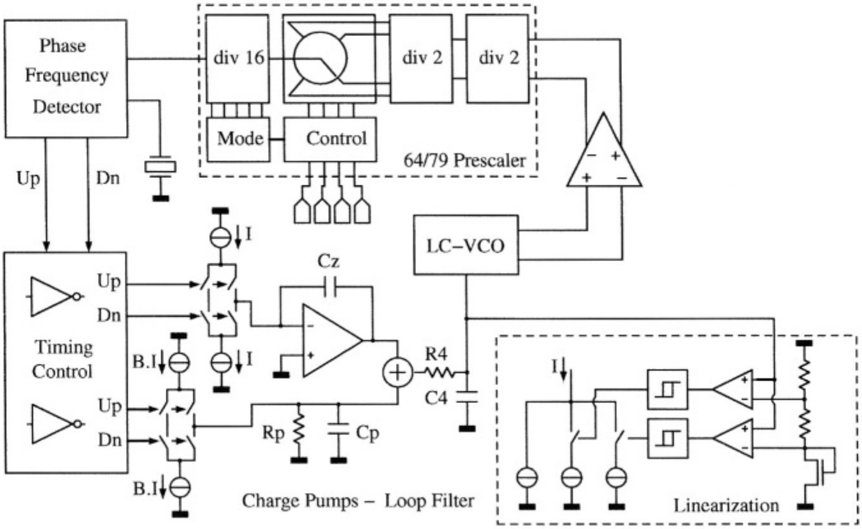
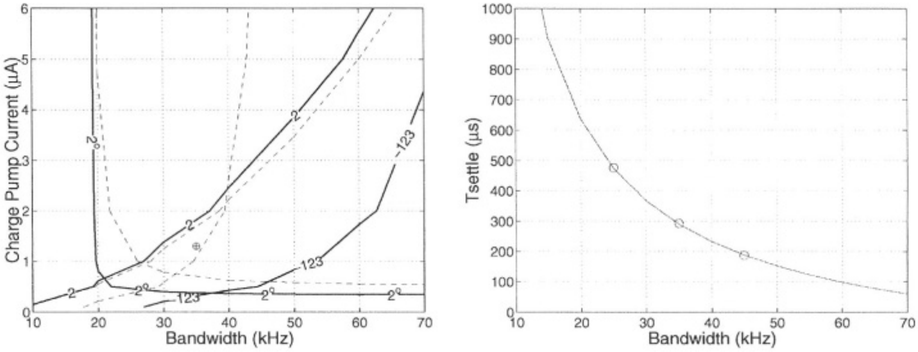

Phase
Frequency
Detector
div 16
div 2
div 2
Mode
Control
64/79 Prescaler
Up
Dn
LC–VCO
Up
Dn
Timing
Control
Up
Dn
I
I
B.I
B.I
Cz
R4
C4
Rp
Cp
Charge Pumps – Loop Filter
I
Linearization

Figure 11: The fully integrated 4-th order, type-II phase-locked loop.



(a) (b)

Figure 12: Matlab simulation results: (a) the design space for N=72 (solid) and N=65 (dashed). (b) The hyperbolically extrapolated loop dynamics for different bandwidths, simulated with Saber.

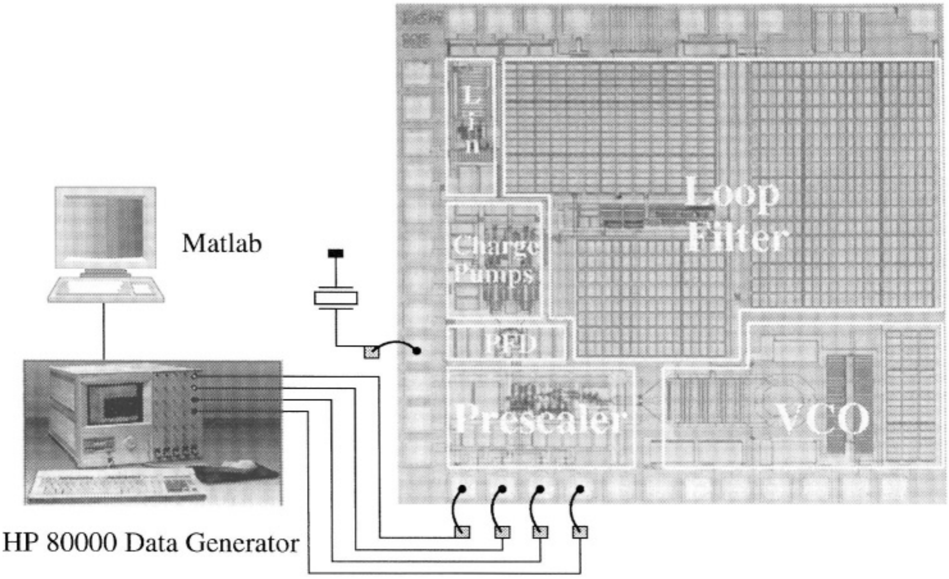
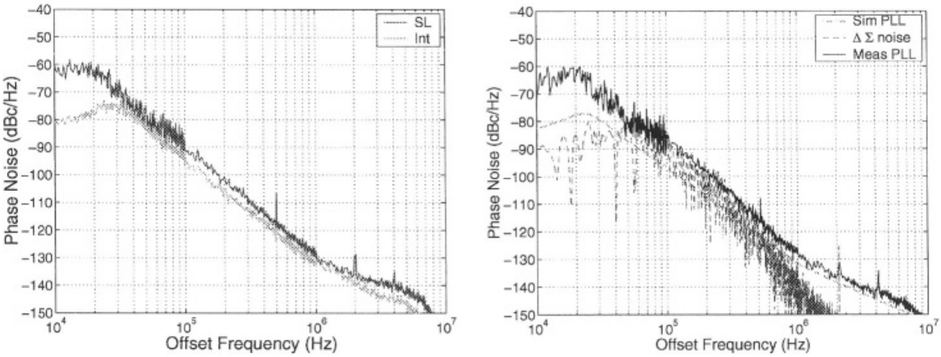


Figure 13: The IC microphotograph and the measurement setup in which it is embedded.

4.3. Experimental Results

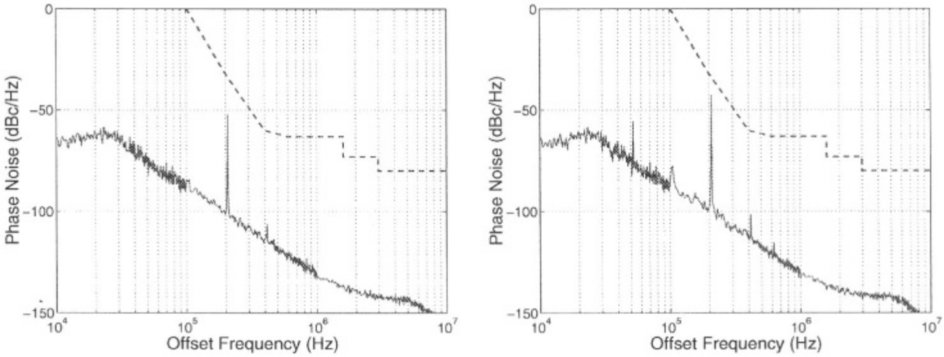
Fig. 13 shows the IC microphotograph and the measurement setup in which it is embedded. The measurements are performed by controlling the PLL divider moduli with a HP80000 data generator, which provides the 4-bit $\Delta\Sigma$ bit-stream generated using Matlab. All presented measurements are performed with a 26 MHz reference frequency and at 1.76592 GHz, i.e. for a fractional division by 67.92. The input to the $\Delta\Sigma$ modulators is a 16-bit word (k=16), resulting in a frequency resolution of around 400 Hz. The power supply voltage is only 2 V.

In Fig. 14 (a), the measured phase noise of the PLL with a multi-bit, single-loop modulator (dark) is compared to the phase noise at integer division (light). Note that the phase noise of the PLL at integer divisions is as low as -124 dBc/Hz at 600 kHz, which is only 0.3 dB higher than predicted by the linear PLL simulations (see Table 1). The measured results for fractional division are much noisier than predicted by the simulations. As a result, the RMS phase error $\Delta\Phi_{rms}$ is increased from 1.7° to 3°. This is due to the high jitter spec of the data generator (> 10ps), that directly translates to in-band noise as confirmed by simulations. This effect is worsened by the limited memory (128kb) of the data generator, which necessitates



(a) (b)

Figure 14: (a) Phase noise measurement of the single-loop modulator compared to the phase noise at integer division. (b) Phase noise measurement with the MASH at 1.76592 GHz compared to the simulated $\Delta\Sigma$ *noise at the output of the PLL (dashed) and with the simulated PLL output without* $\Delta\Sigma$*-control (dash-dotted).*



(a) (b)

Figure 15: Phase noise measurement with the (a) MASH and (b) the single-loop $\Delta\Sigma$ *modulator for a division by 67.9923, which is the worst-case situation in terms of spurious tone suppression. Measurements comply to the DCS-1800 spurious suppression mask (dashed).*

repetition and thus creates large LF spurs that corrupt phase noise measurement calibration at lower offsets.

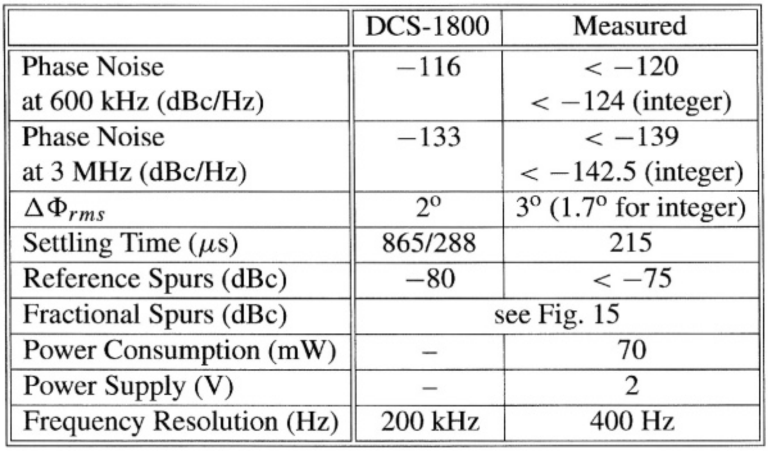
In Fig. 14 (b), the measured phase noise with a MASH converter and the $\Delta\Sigma$ noise as simulated in Section 3. (dashed) is compared. The dashed-dotted line is the simulated phase noise of the PLL without $\Delta\Sigma$-control. The simulated $\Delta\Sigma$ noise leakage closely matches the measured results, except at very low offsets due to the data generator jitter. The spur level is well below -100 dBc, due to careful PFD-charge-pump design. The phase noise at 600 kHz is lower than -120 dBc/Hz, which is close to the simulated values in Table 1. The measured phase noise for the single-loop modulator is however a few dBs lower, as predicted by the simulations (see Table 1).

Measurements have been performed for a division by 67.9923, i.e. 200 kHz from an integer multiple of f_{ref}, which is the worst-case situation in terms of spurious suppression (Fig. 15). The measured fractional spurs at 200 kHz are below -50 dBc for the MASH modulator and somewhat higher for the single-loop modulator with more harmonics, as predicted by the simulations. The measurements are compared to the DCS-1800 spurious suppression mask (dashed) [8, 9]. The spurs are at least 20 dB below the mask for the MASH $\Delta\Sigma$ modulator and 10 dB for the single-loop $\Delta\Sigma$ modulator.

The measured settling time of the PLL is 215 μs for a 95 MHz frequency step. The power consumption of the PLL is 70 mW, 66% of which is due to the monolithic LC-VCO. Table 2 shows the measured specifications compared to the DCS-1800 specifications [8]. The specifications of the IC prototype comply with the DCS-1800, only the $\Delta\Phi_{rms}$ is degraded due to the high jitter of the measurement setup.

5. Conclusions

In this document, the strenghts and weaknesses of $\Delta\Sigma$ fractional-N synthesis have been exposed by means of two analysis approaches. Linear system theory allowed to explore quantitatively the decoupling of the bandwidth and the reference frequency, i.e. the real benefit of fractional-N, given the noise constraints. The analysis revealed that the out-of-band noise specification is the bottleneck. However, because of the inherent switching behavior of a fractional PLL, the linear approach fails. A fast, non-linear analysis method is presented that is able to predict the effects of PLL non-linearities on the spectral purity. Serious in-band noise leakage and re-emerging spurious tones can be observed and are in close correspondence with experimental results. The $\Delta\theta \rightarrow I_{qp}$ conversion building blocks, i.e. the PFD and the charge pumps are identified as the main villains. Using both analyses, a fractional-N PLL design example is elaborated and experimental results are discussed that show compliance with the stringent DCS-1800 cellular specifications.



	DCS-1800	Measured
Phase Noise at 600 kHz (dBc/Hz)	-116	< -120 < -124 (integer)
Phase Noise at 3 MHz (dBc/Hz)	-133	< -139 < -142.5 (integer)
$\Delta\Phi_{rms}$	2^o	3^o (1.7^o for integer)
Settling Time (μs)	865/288	215
Reference Spurs (dBc)	-80	< -75
Fractional Spurs (dBc)	see Fig. 15	
Power Consumption (mW)	–	70
Power Supply (V)	–	2
Frequency Resolution (Hz)	200 kHz	400 Hz

Table 2: Summary of measured specifications compared to the DCS-1800 specifications [8, 9].

All important facts and figures are given, but what can be learned from all this?

MASH or Single-Loop $\Delta\Sigma$ Modulators?

Based on the linear analysis and knowing that a MASH modulator is unconditionally stable and easy to integrate, MASH modulators seem to be the natural choice for monolithic fractional-N synthesizers. However, the presence of a very small non-linearity passes the advantage to single-loop modulators. Due to the shaping of the high-frequency noise by Butterworth poles, the prescaler switching action is less intense. This results in a smaller phase error and thus less noise leakage and less sensitivity to noise coupling, due to the smaller on-time of the charge pumps. In addition, the multi-bit, single-loop modulator offers more flexibility and can easily be made stable by proper choice of the output bits. This comes at the expense of higher fractional spurious tones, but as demonstrated in the example, those are not really critical. Since no design is ideal, single-loop $\Delta\Sigma$ modulators are most promising.

Is CMOS Fractional-N the Way Out?

For low-end applications, the answer to this question is an absolute yes (unless very low in-band noise is mandatory). CMOS fractional-N offers the easiest and cheapest path to integration of low-end systems. The fact that the fractional-N PLL can be re-used as a transmitter even doubles its value [18] (as long as datarates are moderate).

For high-quality transceivers, the answer is less clear and is formulated along the three criteria. Even in the ideal case, the decoupling of the bandwidth and the reference frequency is poor when high spectral purity is the demand. For reference frequencies up to 50 MHz, the maximum loop bandwidth remains well below 200 kHz, i.e. only 1/250. The out-of-band noise spec sets a theoretical limit to the use of the fractional-N technique. In the presence of small PLL non-linearities, significant noise leakage occurs, imposing a practical limit to the rms phase error and phase noise at intermediate frequencies (even at 600 kHz! (see Table 1)). As shown, the re-emerging of spurious $\Delta\Sigma$ tones is less critical in the case of typical cellular standards.

Due to the inherent bandwidth limitation for low noise, the monolithic integration advantage is also endangered. However, for reasonable bandwidths ($\geq$ 35kHz) the loop filter is perfectly integratable, especially when the dual-path loop filter technique is used. This remains true for monolithic solutions with high VCO gains.

Similarly, the agility advantage of the fractional-N PLL can be questioned. However, the given synthesizer example with its low loop bandwidth is still fast enough for the EGPRS multi-slot class 12 ($T_{settle} \leq 310\mu s$ [8]), For fast frequency hopping, special techniques are mandatory anyway [19].

In short, the answer is a silent yes, meaning that fractional-N should always be considered as one of the best options, but it has its limitations…

References

[1] M. Steyaert, J. Janssens, B. De Muer, M. Borrcmans and N. Itoh, "A 2V CMOS cellular transceiver front-end", *IEEE Journal ofSolid-State Circuits,* vol. 35, n. 12, pp. 1895–1907, December 2000.

[2] T. Cho, "A Single-Chip CMOS Direct-Conversion Transceiver for 900MHz Spread Spectrum Digital Cordless Phones", in *ISSCC, Digest of Technical Papers,* February 1999.

[3] A. Roufougaran et al., "A Single-Chip 900 MHz Spread Spectrum Wireless Transceiver in $1\mu\mathrm{m}$ CMOS – part II: Receiver Design", *IEEE Journal of Solid-State Circuits,* vol. 33, n. 4, pp. 547–555, April 1998.

[4] T. Riley, M. Copeland and T. Kwasniewski, "Delta-sigma modulation in fractional-N frequency synthesis", *IEEE Journal of Solid-State Circuits,* vol. 28, n. 5, pp. 553–559, May 1993.

[5] C. A. Kingsford-Smith, *Patent No. 3,928,813,* Washington DC: US Patent Office, 1975.

[6] S. R. Norsworthy, R. Schreier and G. C. Themes, *Delta-Sigma Data Converters: Theory, Design and Simulation,* IEEE Press, 1997.

[7] B. Miller and Robert Conley, "A Multiple Modulator Fractional Divider", *IEEE Transactions on Instrumentations and Measurements,* vol. 40, n. 3, pp. 578–583, June 1991.

[8] B. De Muer and M. Steyaert, *CMOS Fractional-N Synthesizers, Design for High Spectral Purity and Monolithic Integration,* Kluwer Academic Publishers, 2002.

[9] ETSI EN 300 190 (GSM 05.05 version 8.5.1 Release 1999), *Digital cellular communication system (Phase 2+); Radio transmission and reception,* European Telecommunications Standards Institute, 2000.

[10] W. Rhee, Bang-Sup Song and Akbar Ali, "A 1.1-GHz CMOS Fractional-N Frequency Synthesizer with a 3-b Third-Order $\Delta\Sigma$ Modulator ", *IEEE Journal of Solid-State Circuits,* vol. 35, n. 10, pp. 1453–1460, October 2000.

[11] The Mathworks, *Matlab user's guide, version 5,* Prentice Hall, 1997.

[12] B. De Muer and M. Steyaert, *Fully Integrated CMOS Frequency Synthesizers for Wireless Communications,* pp. 287–323, Analog Circuit Design, W. Sansen, J. H. Huijsing, R. J. van dc Plassche (eds.) Kluwer Academic Publishers, 2000.

[13] B. De Muer, M. Borremans, N. Itoh and M. Steyaert, “A 1.8 GHz highly-tunable, low-phase-noise CMOS VCO”, in *Proceedings Custom Integrated Circuits Conference,* pp. 585–588, Orlando, May 2000.

[14] J. Craninckx and M. Steyaert, *Wireless CMOS Frequency Synthesizer Design,* Kluwer Academic Publishers, 1998.

[15] Analogy, *The Saber User Manual,* 1996.

[16] B. De Muer and M. Steyaert, “A CMOS Monolithic Delta-Sigma-controlled Fractional-N Frequency Synthesizer for DCS-1800”, *IEEE Journal of Solid-State Circuits,* vol. 37, n. 7, pp. 835–844, July 2002.

[17] F.M. Gardner, *Phaselock Techniques,* J. Wiley & Sons, New York, USA, 1979.

[18] M. H. Perrott, *Techniques for High Data Rate Modulation and Low Power Operation of Fractional-N Frequency Synthesizers,* Ph.D. Thesis, Massachusetts Institute of Technology (MIT), 1997.

[19] S. Willingham, Michael Perrott, Brian Setterberg, Andrew Grzegorek and Bill McFarland, “An integrated 2.5GHz $\Delta\Sigma$ frequency synthesizer with 5μs settling and 2Mb/s closed loop modulation”, in *ISSCC, Digest of Technical Papers,* pp. 200–201, February 2000.

This page intentionally left blank

A 19mW 2.2GHz FULLY INTEGRATED CMOS SIGMA DELTA FRACTIONAL SYNTHESISER WITH 35Hz FREQUENCY STEP AND QUANTIZATION NOISE COMPENSATION

Ivan Bietti, Guido Albasini, Enrico Temporiti and Rinaldo Castello*
STMicroelectronics, Via Ferrata 1, Pavia, Italy
*Department of Electronics, University of Pavia, Via Ferrata 1, Pavia, Italy

Abstract

A fully integrated 0.18μm CMOS ΣΔ fractional synthesizer targeting 3G wireless terminals applications is presented. This work is a practical example of a more general study on frequency fractional synthesis. A simple linear model of the system is presented and used to simulate different ΣΔ modulators topologies and to evaluate the effects of circuits non-idealities particularly on output spurious tones. Phase Frequency Detector (PFD) and Charge Pump (CP) non-linearity effects are analysed in details, the obtained results are confirmed by measurement. Solutions to overcome these limitations are given. At last, a ΣΔ quantization noise compensation technique is presented allowing to break the trade-off between Phase Locked Loop (PLL) bandwidth and high frequency noise regrowth.

1. Introduction

The last years were characterized by the unrivaled growth of the wireless telecommunication industry. The main technology step that allowed this process was the introduction of digital signal processing in wireless communications, driven by the development of high-performance low-cost CMOS technologies for VLSI. A great concern for integrating a wireless system in a CMOS technology is the challenges associated with implementing a fully integrated high-performance frequency synthesizer. This is reflected in a large effort put into monolithic CMOS integration of RF fractional-N synthesizer both by academics and industry [1] [2].

The goal of this work is the understanding of the fractional synthesizer behaviour and the derivation of methodologies that can drive the choice of the parameters in a fractional Phase Locked Loop (PLL) design.

A. van Roermund et al. (eds.), Analog Circuit Design, 77-96.
©2003 *Kluwer Academic Publishers. Printed in the Netherlands.*

In particular, we applied this knowledge to a full integration in 0.18 μm CMOS technology of a fractional-N frequency synthesizer in order to achieve very-high performances, targeting the UMTS-CDMA standard; anyway this design is not limited to this application, but it can be adapted to many wireless standard only changing the frequency range.

This paper has been organized in the following way: the next section shows the fractional synthesis principles; in section 3 a methodology to investigate the spur generation is explained, and different architectures for the ΣΔ modulator are presented; section 4 presents the key-points in the design of the main building blocks; in section 5 measurement results are shown and section 6 concludes this paper with a further investigation on the generation mechanisms of the spurious tones and on possible techniques to reduce their power.

2. Fractional synthesizers

In a frequency synthesizer based on a PLL architecture, the output is generated by a Voltage Controlled Oscillator (VCO). Its working frequency is set by the "control voltage", i.e. the voltage applied to its control input. This signal is obtained via a negative feed-back loop, comparing the phase of a reference clock (typically generated by an external crystal oscillator) and that of the VCO output after a frequency division by a factor N. The block performing this comparison is called Phase Frequency Detector (PFD), whose output is first filtered and then applied to the VCO input.

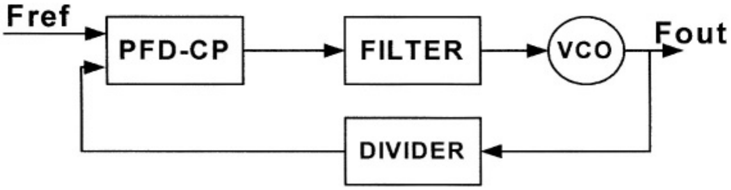


Fig. 1. Schematic diagram of a frequency synthesizer

Provided that the loop is stable, the circuit reaches a steady state condition after a period called "locking time". The negative feed-back ensures that:

$$F_{out} = N \cdot F_{ref} \tag{1}$$

Starting from these first assumptions, it is clear that a PLL can be used as a local oscillator in RF transceivers, since it generates a stable and programmable RF signal, whose frequency is multiple of that of the reference clock. Frequency stability is ensured by slaving the PLL output to the crystal oscillator and programmability is obtained varying the division ratio N in the feedback path.

Classical integer PLL have some limitations that do not allow the wide use in many RF systems. The strongest one is the limited frequency resolution of the output frequency which is forced to be a multiple of Fref. Combining this with the need to perform channel selection at RF (which is very often the case in super-heterodyne transceiver and it is mandatory in direct conversion architecture) forces to adopt very low reference frequency. The situation is further aggravated if the synthesizer has to be used in a multi standard transceiver, where the frequency plan is very fragmented and the resulting channel spacing can be very narrow. A very low Fref has many drawbacks: it requires an high value for N, so the phase noise contribution of the divider, the phase comparator and the input signal is strongly emphasized (its power at the output is multiplied by the factor N^2). Moreover, since the loop bandwidth of the PLL must be much lower than Fref for stability reasons, this causes an often unacceptable long locking time. Furthermore, noise contribution of different blocks is not optimized.

Fractional PLLs can overcome these limitations thanks to the non-integer oscillator frequency division capability. For example varying the division ratio between the integers N and N+1, results in an average fractional division value which is intermediate between them.

In this case, in lock condition,

$$F_{out} = (N + K/M) \cdot F_{ref} \tag{2}$$

where K is number corresponding to the selected channel frequency and M is the fractionality depth, i.e. the total number of selectable output frequencies.

Among the benefits of this architecture there are: fast locking time, agile frequency switching, potentially arbitrary output frequency resolution, and more freedom in the reference frequency choice.

In our design the fine frequency resolution capability has been fully exploited in order to achieve 35Hz step; our target is to make the synthesizer able to compensate crystal frequency drift with a digital word, avoiding the need for an analog compensation using an expensive Voltage Controlled Crystal Oscillator.

The main drawback of this fractional technique is the generation of spurious tones around the output carrier. They are at frequency offsets that are harmonics of the inverse of the periodicity in the division control pattern. This pattern, whose averaged value corresponds to the fractional number in the division factor, can be generated using many techniques, the most popular being a ΣΔ modulator.

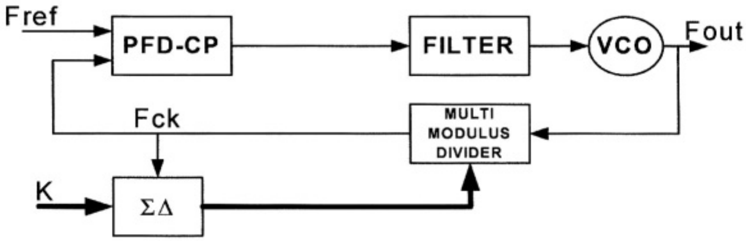


Fig.2. Schematic diagram of a fractional synthesizer

ΣΔ modulators are used because of their capability in spreading and shaping the quantization noise [3]. In this application, the equivalent of the quantization noise in the PLL is the phase error, i.e. the difference between the phase due to an ideal fractional division and the one due to the actual division selected by the pattern. This error is not random, but has a periodicity depending on the modulator type, the fractionality depth and the selected frequency. As a consequence, the unwanted signal at the output is not random phase noise, but a series of spurs at the harmonics of the frequency corresponding to the ΣΔ periodicity. The power of these harmonics is shaped by the ΣΔ modulator moving most of their power at high frequency, in similar way as the quantization noise in the classical ΣΔ converter.

The advantage of this noise shaping is that the great majority of the spur energy can be attenuated by the loop filter. In fact the phase transfer function from the divider to the VCO output is low-pass type. The higher the order of the ΣΔ modulator, the more effective the shaping to high frequencies of the spurs; but also the higher the required order of the Low Pass Filter (LPF) to filter them out. The order of the ΣΔ modulator must be equal or lower than the high frequency slope of the loop transfer function. This is because a first order filtering effect on the spectrum of the ΣΔ quantization noise is implicit in the frequency to phase conversion performed by the divider [4]. In practice a maximum order filter exists since high orders degrade the stability of the synthesizer loop.

A very important parameter is the bandwidth of the loop. As previously described, the cut-off frequency should be as low as possible, in order to filter the shaped spurs. In fact the residual unfiltered spur power depends on the order of the modulator and on the ratio Fref/bandwidth in the same way as over-sampling ratio in the ΣΔ converter. Lower bandwidth is also an advantage in order to filter other spurious signals (e.g. at Fret) and the phase noise generated by the divider, the phase comparator and the loop filter. On the contrary, high bandwidth is requested in order to filter out the phase noise of the VCO and to achieve a fast settling time, usually a stringent specification in current standards. As rule of thumb, locking time can be evaluated as five times the inverse of the bandwidth

(in Hertz).
Moreover large bandwidth fractional synthesizers can be used in direct modulation architecture. In fact in this case the fractionality value K can be modulated by a signal s(t), so that the output frequency is

$$F_{out} = [N + K/M + s(t)/M] \cdot F_{ref} \quad \textbf{(3)}$$

provided that the bandwidth of the loop is large enough to avoid filtering [5].
In this design we designed the 3dB closed loop bandwidth to be 600kHz. This gives very fast settling time and allows direct modulation of the carrier for some wireless standard.

3. Fractional synthesizer simulation

A fractional PLL is a relatively complex system. Therefore a complete transistor level simulation requires a too long computational time. A solution to obtain accurate results consists in the realization of a behavioural model (e.g. in a VerilogA/MATLAB environment) where each block is replaced by its time-domain description, thus allowing the closed loop simulation of the structure. The resulting time domain model is much more simple than the transistor level structure; however, a quite long computational time is still needed to simulate channels characterized by a long ΣΔ limit cycle (more than one day of CPU time on a SPARC 60 in case of a few Hertz frequency resolution).
A further reduction in complexity can be obtained if a frequency-domain linear time invariant (LTI) model is introduced [6]. In such a model, the intrinsically time variant divider is replaced by a time invariant divider, with a non integer dividing ratio whose value depends on the selected channel. The remaining blocks are replaced with their own linear model. However, also the high frequency shaped quantization noise injected by the ΣΔ in the feedback path of the loop should be accounted for. This can be approximately obtained in the way shown in Figure 3.

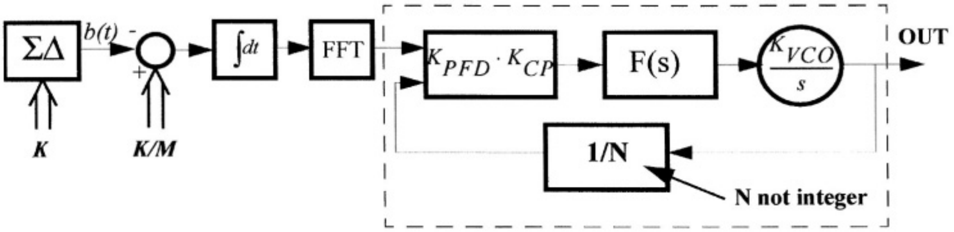


Fig.3. Frequency-domain LTI model for spurs simulation

The ΣΔ output sequence b(t), corresponding to the selected channel, no longer modulates the division factor of the divider, but it is passed through an integrating block to convert it into phase error, and its frequency spectrum is sent to the reference input of the PLL. For further simplification, the whole PLL, dashed box of Figure 3, can then be represented by its equivalent closed loop transfer function. The outcome of this simulation represents the PLL output spectrum referred to the carrier. It is worth saying that this approach is valid if the filtering effect of the LPF sufficiently reduces the high frequency portion of the ΣΔ quantization noise. In other words this technique is equivalent to open the loop at the output of the VCO (whose spectrum is what we are interested in) and substitute it with a perfectly clean clock at the input of the fractional divider. Unfortunately, in the real case, part of the quantization noise is not filtered out by the LPF causing a second order effect which is not taken into account by the model.

The comparison between the results of the slow but accurate time-domain model and of the approximated but fast frequency-domain linear model are reported in Figure 4 for channel 451584 assuming a third order ΣΔ modulator with 2^{20} possible channels. The difference between the two is negligible.

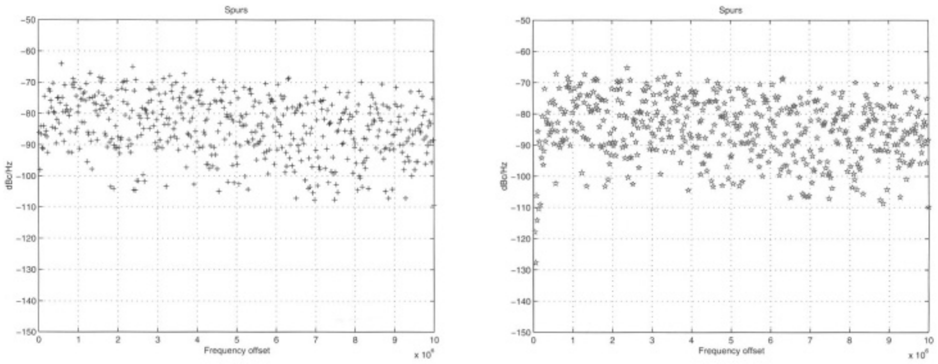


Fig. 4. Comparison between time-domain model (left) and frequency-domain linear model (rigth) for channel 451584/2^{20}

Therefore most of the following simulations results have been obtained using the frequency domain model, that requires a reasonable computational time also for channels characterized by a long ΣΔ periodicity (less than one hour CPU time on a SPARC 60 MATLAB simulation for the longest one).

Such a frequency domain model can be easily modified to keep into account also non linearities of the Phase Frequency Detector + Charge Pump (PFD + CP). These effects can be reasonably modelled passing the time domain signal at the output of the integrator through a non linear block before feeding it into the FFT function. Let us consider, for example, the CP non idealities: positive and negative slopes of the PFD+CP I/O characteristic (i.e. charge injected into the loop filter *vs* input phase difference) are quite constant (<0.1% variation); however their value is generally different between positive and negative input phases. This causes a sharp non linearity of the transfer function around zero, whose effect is quantified in Figure 5.

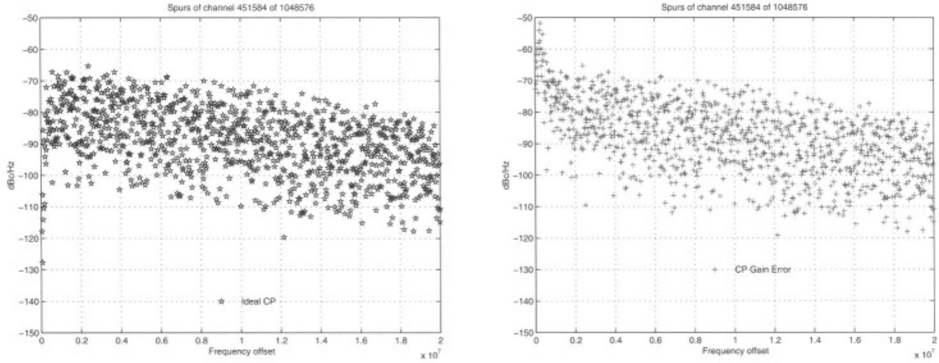


Fig.5. Output spurs level for channel 451584/2^{20}: frequency-domain linear model (left); frequency-domain model with ±1% non linearities (right)

The left graph represents the ideal output spurs due to ΣΔ quantization noise for channel 451584, while the right one shows the output spurs in presence of ±1% slope difference. The effect of non linearities is a strong increase of in band spurs. Particularly, a large amount of power concentrates at the so-called "fractional frequency", corresponding to $(K/M) \cdot F_{ref}$. The amount of this effect will be quantified in section 5.

To characterize the actual amount of non idealities associated to the single blocks, e.g. PFD+CP non linearities or VCO phase noise, transistor level simulation is necessary. This will be described in detail in section 4.

The frequency-domain model (LTI) has been very useful in order to compare different possible architecture for the ΣΔ modulator. In fact many solutions exist, e.g. single-loop, multi-loop or MASH, with different order and number of output bit [3]. Moreover a simulation over every allowed fractional division is necessary in order to check the spur performances for each synthesized channel, since the PLL output spectrum is very different varying the periodicity of the pattern. This study focused on two parameters: The level of the higher spur and the total integrated spur power over the frequency range [1kHz - 1.94MHz]. The simplest implementation is the first order single bit ΣΔ modulator, equivalent to an accumulator, which has a maximum periodicity equal to the fractionality depth M. In this case either the highest spurs level is unacceptable (in the case of low periodicity) or the total integrated spur power is too high (in the case of many spurs in the frequency range due to long periodicity). Other solutions with multi-loop architectures require more adders; in this case the spectrum is better shaped and the periodicity is longer, at the cost of more power burned in the modulator. Modulators with multi-bit output show even better results in both the considered performance parameters, but they need a multi-modulus divider, instead of a dual-modulus one; this building block is more complex (usually the "swallow architecture" is adopted), and requires burning more power. Among the multi-bit architectures, the most popular implementation is the MASH one. This can be implemented cascading single bit modulators with different order (e.g. MASH 1-1, 2-1, 1-1-1); the outputs of each block are combined to generate the modulator output.

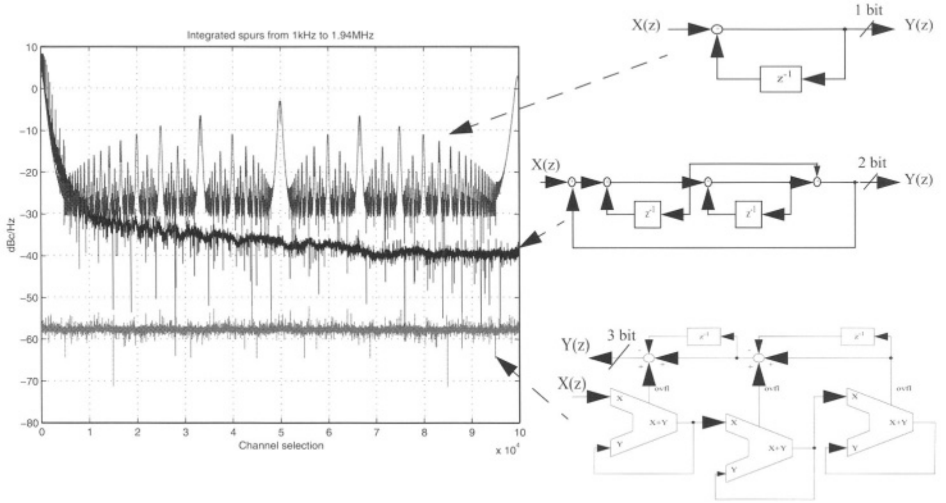


Fig.6. Spur power with different ΣΔ architecture for all output frequency

After extensive simulations for different architectures over all the possible channels, the choice fell on the MASH 1-1-1 (see Figure 6); to achieve the targeted 35Hz resolution with a 35MHz Fref, a 20 bit fractionality depth was chosen. The loop filter has been designed to have a 3dB closed loop bandwidth of 600kHz and a third order high frequency slope (one higher than the phase quantization noise produced by the modulator).

4. PLL Building Blocks Design

The realization of a fractional synthesizer is quite challenging since it makes use of several different types of analog circuits. In fact VCO design faces classical RF problems, multi-modulus dividers requires high speed circuitry knowledge and for charge pump and filter implementation a good experience on precision low frequency design is needed. Finally some knowledge on digital synthesis is useful for the $\Sigma\Delta$ integration and experience on avoiding cross-talk between blocks is mandatory. This section will explain how all the previously mentioned building blocks have been designed in this work.

VCO: The best performance in terms of phase noise vs power are obtained using LC tank oscillators. This topology is generally used in the gigahertz range (which is our case) since for lower frequency the inductor value becomes to high to be integrated. Its working principle is relatively simple: an active negative resistor (generally built using a cross coupled differential pair) is added to an inductor and a capacitor tank in order to compensate its losses and sustain a stable oscillation. The spectral purity of this oscillation signal is mainly defined by the quality factor of the tank which, at 2GHz, is dominated by the inductor losses. The loss amount is directly related to the resistivity of the metal used for the spiral and to the substrate conductivity. In our case a $10\Omega cm$ substrate and a combination of metal layers giving a total of $20m\Omega_{\square}$ resistance have been used, allowing an 'on-chip' inductor quality factor Q around 9 @ 2.1GHz. Adding the variable capacitor losses, a total Q of 8 has been measured.

Considering that the best noise performance is reached when the amplitude of the oscillation is maximized (i.e. $1.8V_{pk}$ in the technology used) and assuming the above Q value, the theoretical minimum thermal phase noise using 1.5mA current (as in the integrated VCO) is -128dBc/Hz @ 1MHz offset from the 2.1GHz carrier [7]. This calculation has been done assuming that, at the resonance frequency, the bias current is completely switched into the resonant load and filtered by the load itself and that the only noise source is the loss associated to the tank.

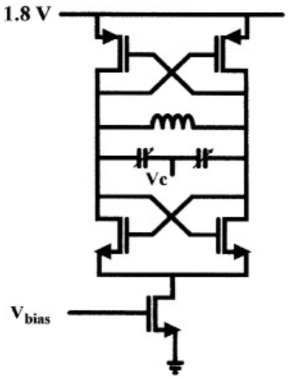


Fig. 7. VCO circuit

The implemented circuit, see Figure 7, has some unavoidable non-idealities: first the negative resistor is noisy and second, part of the available oscillation swing is lost due to the voltage drop on the current tail and on the cross coupled pairs. In our design starting from a 1.8V supply a peak oscillation amplitude of 1.2V is achieved. The implemented negative resistance has been chosen to be half the value of the tank losses. Under these new conditions the minimum phase noise @ 1MHz offset results to be -121dBc/Hz which is very close to the measured -119.5dBc/Hz leaving a small design margin to improve the VCO performance. Responsible for this difference are the tail generator noise and devices flicker noise up conversion.

It must also be noticed that the use of a complementary (PMOS NMOS) structure has some advantages. First of all it allows the use of the standard supply voltage: in fact to achieve the same performance in a classical topology burning the same power a second supply equal to approximately half the maximum swing is needed. Second the more symmetric structure reduces the up conversion of the switching pairs flicker noise [8].

Regarding the frequency control, an MOS varactor has been used. This device is available in all the new standard CMOS technologies and it consists of an NMOS built into an NWELL [9]. Its capacitance variation is approximately 1:3 while its quality factor is higher than 30 @ 2GHz.

LPF: The NWELL back-plate of the varactor is directly tied to the loop low pass filter output. As shown in Figure 8 a passive structure has been preferred to an active one.

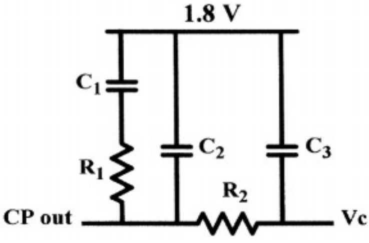


Fig.8. Loop filter circuit

The reason of this choice, paid with a larger area, is to improve immunity from supply or substrate noise present on chip due to the ΣΔ digital circuitry. In fact the loop filter is connected between frequency control node and the positive power supply to which the oscillator common mode voltage is referred. The use of an N-type varactor and of an oscillator common mode referred to the positive power supply is the main source of noise injection from the supply to the voltage control of the VCO. In fact the varactor NWELL to P-substrate parasitic capacitance together with the LPF C_3 capacitor creates a capacitive voltage divider between the supply and the control node of the VCO. This drawback can be solved using a version of the oscillator complementary to the one described here (so with the common mode referred to ground) while keeping the same varactor. This has not been done in the present implementation because the chosen solution allows a wider variable capacitor tuning range.

Phase Frequency Detector + Charge Pump: The PFD and the CP used in this work are relatively standard, but particular care has been put on linearizing their response. In fact, as it has been pointed out in section 3, a non linear characteristic of the output charge into the loop filter versus the input phase error, strongly influences the performance of a not-integer synthesizer particularly on the fractional spur amplitude. This problem is negligible in classical integer N synthesizers.

Two kind of non linearities have been studied: the already mentioned asymmetric response to negative and positive phase errors and the non linearity around zero. It has been observed that the first type is dominated by the charge pump performance, due to the positive and negative current source matching. The implemented circuit is shown in Figure 9.

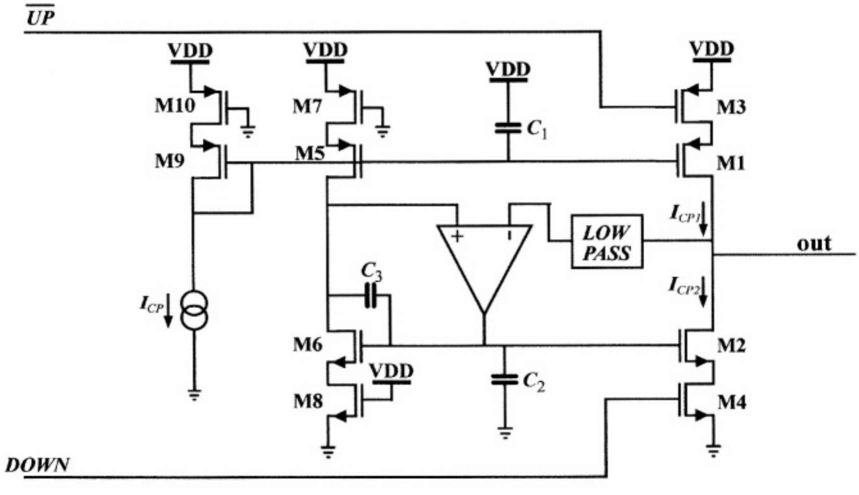


Fig.9. Charge Pump circuit

Any transistor drain resistance effect has been drastically reduced using a feedback loop which forces the drain voltage of a replica pairs of transistors to be the same as that of the P and N current sources. The only remaining cause of P-N asymmetry is the mismatch between the replica devices and the current generators. Although not done in the current implementation it is possible to use an auto-zero technique running at the start up, in order to inject a DC current into the feedback loop and compensate for the mismatch.

The non linearities around the zero phase error are mostly due to the phase frequency detector behaviour e.g. the existing dead zone of the PFD. As usually done this problem has been solved introducing a delay in the two flip-flops reset path (see Figure 10a).

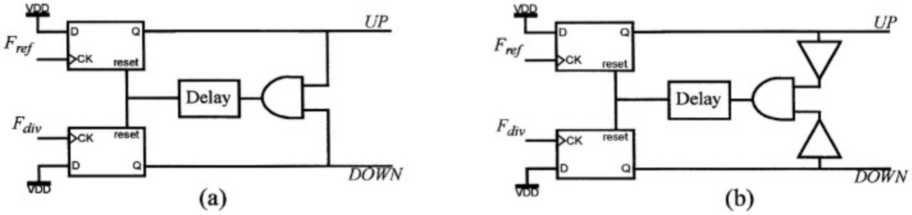


Fig. 10. PFD implementations

However a more subtle problem causing a gain enhancement around zero still remain. In fact when input phases are very close each other, the load presented by a CMOS NAND gate changes (i.e. becomes lower) compared to the case when they are far away. This causes a small variation in the length of the Up and Down pulses sent to the CP. The problem can be easily solved introducing a buffer as shown in Figure 10b. Quantification of this phenomenon will be discussed in section 5.

A multi-modulus prescaler is needed because a 3 bit MASH 1-1-1 ΣΔ has been chosen. It has been built using the swallow counter technique (Figure 11) and a dual-modulus divider capable to count by 3 and by 4.

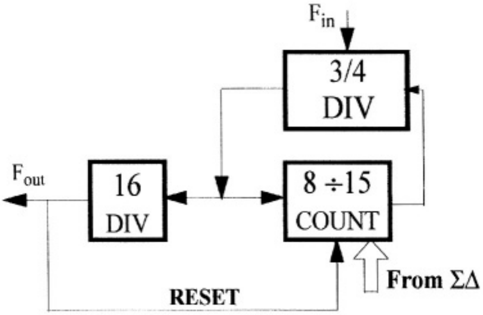


Fig. 11. Swallow counter architecture

The choice of this two modulus is very efficient since it can be implemented using 3 synchronous flip-flops with the possibility to close only 2 of them in a negative loop (division by 4) or all them in a positive one (division by 3). Each time the ΣΔ calculates a new division ratio, a total number of 16 divisions (part of them by 3 ad part of them by 4) is performed. The result is that a programmable number between 56 and 63 of VCO cycles is counted before generating the phase edge to send to the PFD.

The logic style for this digital circuit is part current mode (CML), part pure CMOS. In particular the high frequency prescaler has been designed using pseudo differential CML where only NMOS devices appear in the signal path (see Flip Flop latch in Figure 12).

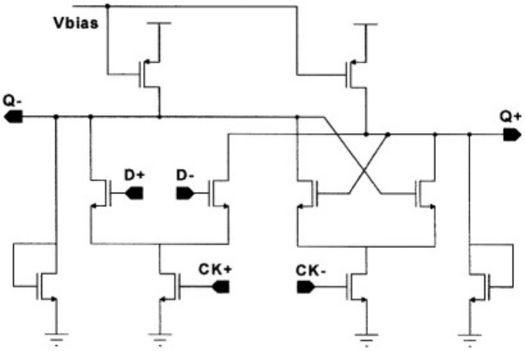


Fig.12. Latch circuit

To complete the full integration of the fractional synthesizer a digital MASH 1-1-1 **ΣΔ modulator** with three 20 bits accumulators has been integrated on chip. This is not very common since many of the published works make use of external patterns generators. In fact a significant amount of digital circuitry running at a frequency which is not an integer sub-multiple of the VCO output can potentially create cross-talk, and therefore fractional spurs, problems. For this reason all the digital circuitry has been isolated using a triple well structure consisting of an N-type tub containing the digital devices. Dedicated digital power supply have been also used. To further improve immunity, the timings of the ΣΔ switching and of the phase comparison have been spaced apart. In this way the charge spikes injected into the substrate at each digital edge are allowed to decrease before phase comparison occurs.

5. Test chip measurements

A fully integrated version of previously presented circuit has been realized in a 0.18μm CMOS technology. A reference oscillator is the only external component required to realize the complete fractional synthesizer. A dual bias voltage (1.8V and 3.3V) has been used, with the high voltage needed essentially to extend the tuning range of the varactor. 0.8mA are sunk from the 3.3V supply and 8mA from the 1.8V supply, corresponding to a 17mW total power consumption. A photograph of the integrated prototype is reported in Figure 13: The chip size is 1.9mm x 1.6mm including pads.

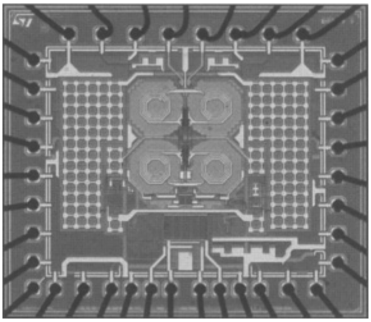


Fig. 13. Photograph of the fully integrated prototype in 0.18μm CMOS

The test chip has been used to verify the achievable level of performance and to validate the accuracy of the introduced model.

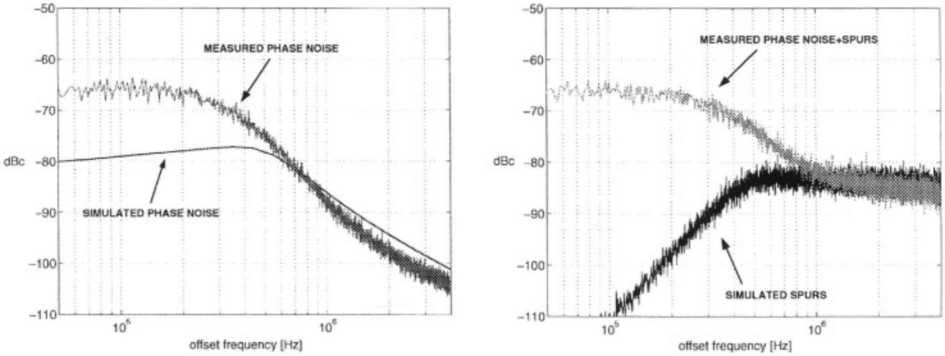


Fig. 14. Output phase noise and spurs (RBW=1kHz): transistor level simulation vs measurement of phase noise for ch. 0/2^{20} (left); frequency domain simulation vs measurement of phase noise+spurs for ch. 1/2^{20} (right)

Figure 14 left shows the transistor level simulation results and the measurement for the PLL phase noise, corresponding to the selection of channel 0, i.e. a situation in which the $\Sigma\Delta$ is not working and the fractional PLL is equivalent to an integer one (allowing a transistor level simulation in a reasonable time). Both curves are obtained making use of a resolution bandwidth of 1kHz. Good agreement between simulations and measurement is obtained at frequency offsets higher than 600kHz. At lower frequencies the measured phase noise is dominated by external parasitic effects. In Figure 14 right is represented the result of the previously described frequency-domain linear simulation of the spurs corresponding to the selection of channel 1, a situation in which the $\Sigma\Delta$ is working with the LSB asserted in order to randomize spurs power. Resulting spurs are 17.5Hz spaced from each other; 30dB/dec slope at low frequency offsets is visible, due to 3rd order $\Sigma\Delta$ modulator shaping. The curve is obtained using a resolution bandwidth of 1kHz. Corresponding measured phase noise + spurs level is represented in the same figure. Very good agreement with previous results is shown. The fractional spur level is at a 35Hz frequency offset from the carrier and therefore too close to the carrier to be detected in the measurement. A dual situation, in which the fractional spur cannot be detected in the measurement, consists in the selection of a channel characterized by a fractional frequency much higher than the loop bandwidth. Consider for example channel $2^{19}+1$, which is randomized through the assertion of the LSB not to have high output spurs due to $\Sigma\Delta$ quantization error. This channel is thus characterized by the long $\Sigma\Delta$ periodicity of 1/17.5Hz (corresponding to a 17.5Hz spacing between spurs) and a fractional frequency of 17.5MHz. No tones are visible at the fractional frequency due to heavy attenuation by the loop filter. In order to present the increase of the fractional spur level due to the non linearity, Figure 15 shows the measured phase noise + spurs level for the randomized channel $2^{17}+1$ with a resolution bandwidth of 10kHz.

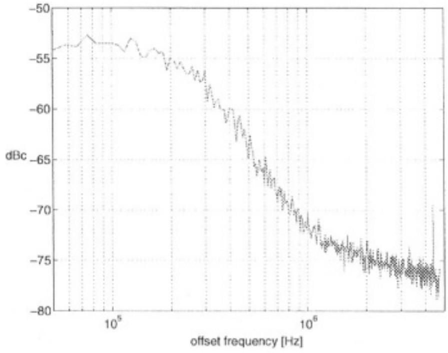


Fig.15. Measurement of phase noise + spurs for ch.$(2^{17}+1)/2^{20}$
(RBW=10kHz)

Channel $2^{17}+1$ is characterized by a $\Sigma\Delta$ periodicity of 1/17.5Hz and a fractional frequency of 4.4MHz. A sharp tone is present at the fractional frequency, corresponding to non linear effects in the PFD+CP and to intermodulation effects. The fractional spur level is quite small (~-70dBc) because the examined fractional frequency is higher than the loop bandwidth, and power at the fractional frequency is thus partially attenuated by the loop filter. A frequently used trick in order to reduce fractional spur level consists in lowering the loop bandwidth, thus attenuating fractional spur level. Anyway, such a trick does not intrinsically solve the problem, since it is still possible to choose a proper set of bad channels whose fractional frequencies are lower than loop bandwidth, and thus not attenuated.

In order to investigate fractional spur generation mechanisms a second version of the same circuit was realized with external loop filter and VCO. Some features were added to the chip, that allow the operation of single building blocks. In particular PFD+CP were deeply investigated when operating in an open loop mode. In order to appreciate the small non linearity of PFD+CP I/O characteristic, the derivative of such a characteristic was plotted. This derivative can be easily obtained by measuring the small signal demodulation gain of the PFD+CP.

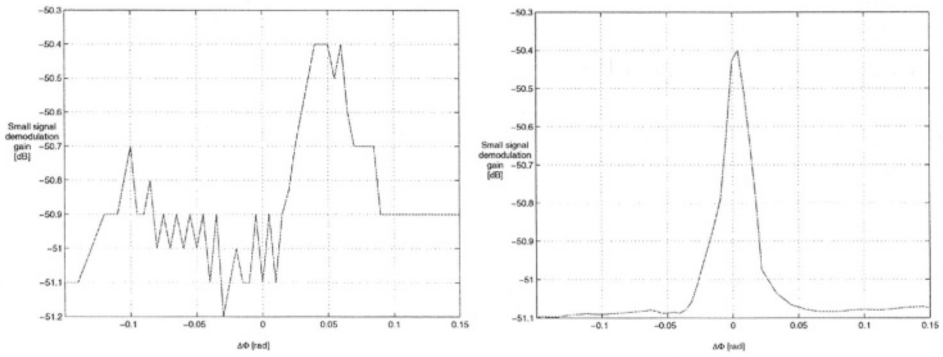


Fig. 16. Derivative of PFD+CP I/O characteristic: measured small signal demodulation gain (left); transistor level simulated small signal demodulation gain (right)

In the left curve of Figure 16 the result of such a measurement is plotted; a ~0.6dB increase in the value of the slope at small ΔΦ values is shown. Results very similar to measurement are shown by a transistor level simulation of the block (right curve in Figure 16). This non linearity, introduced in the previously described frequency-domain simulation for the spurs, causes an increase in low frequency spurs level and particularly in fractional spur level, as shown in Figure 17 for channel 257, whose fractional frequency falls inside the loop bandwidth.

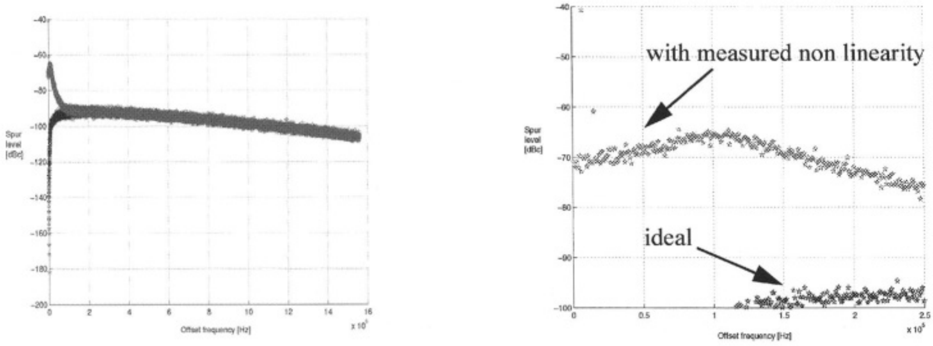


Fig.17. Increase of low frequency and fractional spurs due to nonlinearities for ch. 257/2^{20} (left=full scale; right=zoom)

Such a level is very close to the measured value of ~-40dBc. The main cause of fractional spur increase was then identified in PFD+CP I/O non linearities.

To reduce the effects of these non linearities a DC current was injected into the loop filter (where also the CP injects its current). This causes the PLL to work, in locking conditions, around a **ΔΦ** value different from 0, i.e. on a more linear part of the characteristic. In fig.10 the measured level of fractional spur for channel 257 (with in band fractional frequency) is plotted, as a function of injected DC current.

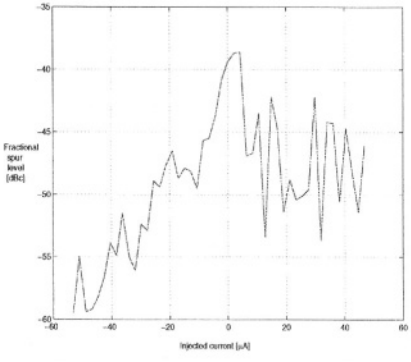


Fig.18. Measured fractional spur level for different values of DC injected current for ch. 257/2^{20}

The figure shows that for large enough sunk current the fractional spur level can be decreased from ~-40dBc to ~-55dBc. Such an implementation has the disadvantage of carrying a ripple at the reference frequency on the control voltage, which produces an increase of the output reference spur.

More sophisticated circuit linearization techniques are currently under test, together with architectural techniques in order to reduce non linearities effects, without degrading neither reference spur level nor phase noise.

6. Analysis of spurs generation mechanism

As it has been remarked in the whole paper, the understanding of the spurs generation mechanism is the key point of a fractional synthesizer design. The causes of spurs can be grouped into three categories: quantization noise associated with the intrinsic ΣΔ structure, blocks non linearities (as explained in the sections 3 and 5), analog crosstalk. The first one is totally predictable since it comes from a known sequence generated by the modulator itself and, in principle, can be compensated for. In practice it is possible to digitally calculate a sequence vector representing the charge injected into the loop filter by the charge pump (responsible for the ΣΔ spur generation) and to use it to drive a digital to analog converter (DAC) for compensation. This concept is represented in Figure 19.

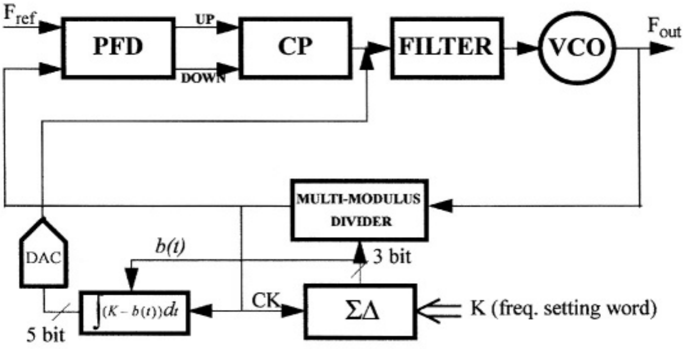


Fig.19. Technique to compensate the spurs due to quantization error

Unfortunately the principle works only if the DAC resolution is the same of the ΣΔ accumulator length (i.e. 20 bits) and the same for its linearity. If the DAC performance arc much lower, the effect is that the compensation is effective only for the very high frequency quantization noise while, for the frequencies close to the carrier, things get worse. This has been proved integrating a 5 bit DAC converting the truncated ideal 20 bit word. The measurements are reported in Figure 20.

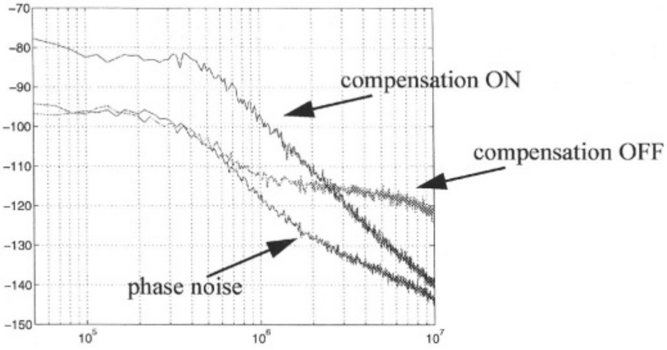


Fig.20. Measured results of the compensation by a DAC with reduced number of bits

At an offset frequency from the carrier above 2MHz an improvement is obtained, while for lower frequency offsets this technique produces a strong degradation. The way to overcome these limitations has been identified and proved with simulations. It consists of three points: high pass filtering of the compensation current, DAC elements scrambling to spread over a large band the energy produced by its limited linearity, DAC quantization error shaping via another ΣΔ algorithm. All these solutions have been implemented in a test-chip which is currently under testing.

Particular attention has to be paid to the inter-blocks crosstalk. This is an intrinsic problem of fractional PLLs because not multiple frequencies are simultaneously present on chip. Any non linear phenomenon (intermodulation, sampling etc.) can produce a phase modulation of the VCO output. This is particularly evident if part of the RF signal is coupled to the PFD-CP supply. Since these blocks do not have infinite PSRR and they are intrinsically not linear form the power supply, a signal at the fractional frequency will appear at their output (see Figure 21).

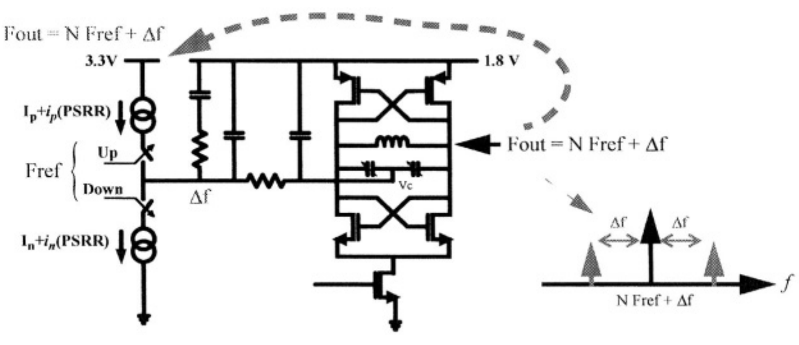


Fig.21. Coupling example of output signal on CP supply

The effect of the LPF on this signal can be negligible if the intermodulation frequency is low, thus the VCO will be modulated. Other examples of the same kind can be easily identified. Quantifying this phenomenon is also a problem. In fact, since all of them produce a spur at the fractional frequency, they can sum or compensate depending on the relative phase. For this reason a second 'split blocks' chip has been integrated as mentioned in section 5. The measurements have shown an amount of cross-talk equivalent to a -50dBc in band spur. We are confident to reach -60dBc in the next version improving the isolation techniques but our feeling is that better values will be hard to achieve.

References

[1] Chi-Wa Lo et al., "A 1.5-V 900-MHz Monolithic CMOS Fast-Switching Frequency Synthesizer for Wireless Applications", IEEE Journal of Solid State Circuits, vol. 37 n° 4, April 2002

[2] W. Rhee et al., "An 18-mW 2.5-GHz/900-MHz BiCMOS Dual Frequency Synthesizer with <10-Hz RF Carrier Resolution", IEEE Journal of Solid State Circuits, vol. 37 n° 4, April 2002

[3] T.A.D. Riley et al., "Delta-Sigma Modulation in Fractional-N Frequency Synthesis", IEEE Journal of Solid State Circuits, vol. 28 n°5, May 1993

[4] A. Marques et al., "Theory of PLL fractional-N frequency synthesizers", Wireless Networks, vol. 4 issue 1, January 1998

[5] N.M. Filiol et al., "An Agile ISM Band Frequency Synthesizer with Build-in GMSK Data Modulation", IEEE Journal of Solid State Circuits, vol. 3 n° 7, July 1998

[6] M.H. Perrott et al., "A Modelling Approach for $\Sigma-\Delta$ Fractional-N Frequency Synthesizers Allowing Straightforward Noise Analysis", IEEE Journal of Solid State Circuits, vol. 37 n° 8, August 2002

[7] T. Lee et al., "Oscillator Phase Noise: A Tutorial", IEEE Journal of Solid State Circuits, vol. 35 n° 3, March 2000

[8] A. Hajimiri et al., "A General Theory of Phase Noise in Electrical Oscillators", IEEE Journal of Solid State Circuits, vol. 33 n° 2, February 1998

[9] F. Svelto et al., "A 1.3 GHz Low-Phase Noise Fully Tunable CMOS LC VCO", IEEE Journal of Solid State Circuits, vol. 35 n° 3, March 2000

IMPLEMENTATION ASPECTS OF FRACTIONAL-N TECHNIQUES IN CELLULAR HANDSETS

Yann Le Guillou, Dominique Brunel
Philips Semiconductors
2 rue de la girafe B.P.5120, 14079 Caen Cedex 5, France

Abstract

High performance fractional-N synthesizers based on **ΔΣ** modulation have significantly increased the level of integration in wireless handset transceivers. [1]
This paper describes the operation of **ΔΣ** fractional-N PLL and discusses performance impairments such as tones, PFD-charge-pump non-linearity and mismatch. Practical guidance will be given to define a **ΔΣ** topology. Finally, examples of fractional-N synthesis techniques used to provide LO to a receiver IQ mixer, UMTS clock and variable IF in transmitter will illustrate the necessity of this architecture in the RF system portfolio [1].

1. Introduction

The phase-locked loop (PLL) synthesizers are the most popular of all frequency synthesis techniques and are used in all analogue RF parts in wireless transceivers [2]. Since the last decade, PLL techniques have been improved using **ΔΣ** fractional-N PLL that allows digital shaping of noise and spurious signal. This paper first describes the basic operation of such a synthesizer that uses **ΔΣ** technique. In section 3 **ΔΣ** modulator topologies and implementations as well as phase/frequency, charge-pumps and tones impairments are discussed. Finally, section 4 illustrates the extensive uses of **ΔΣ** fractional-N PLL in the exciting area of multi-mode RF systems.

2. ΔΣ fractional-N overview

ΔΣ Fractional-N PLLs are built with conventional synthesizers components such as voltage-control oscillator (VCO), phase/frequency detector (PFD), charge-pump (CP), loop filter (LF) programmable dividers and with a **ΔΣ** modulator (Fig. 1).

A. van Roermund et al. (eds.), Analog Circuit Design, 97-110.
© 2003 *Kluwer Academic Publishers. Printed in the Netherlands.*

This modulator was first used in A/D and D/A converter [3]. The fundamental operation in ΔΣ modulators is that the quantization noise spectrum is shaped in such a way that small amount of noise power remains in the useful signal bandwidth. The rest of quantization noise is pushed to high frequencies with a slope directly related to the modulator. The same principle is applied in ΔΣ Fractional-N synthesis by pushing the phase error toward high frequencies so that the phase noise close to carrier is small [4]. The high frequency phase error is then suppressed by a low-pass loop-filter.

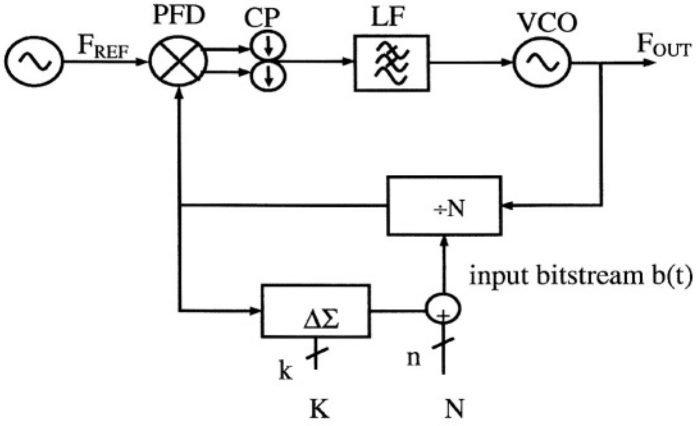


Fig .1: ΔΣfractional-N PLL architecture

3. Architecture trade-offs

3.1. The ΔΣ modulator

A 1st order ΔΣ modulator generates high level of spurii and its flat or non-shaped phase noise spectrum at PLL output degrades the noise close to carrier frequency. Consequently, higher order modulators have to be used. Usually we can distinguish two types: the single loop or feedback type and the cascaded or MASH type.

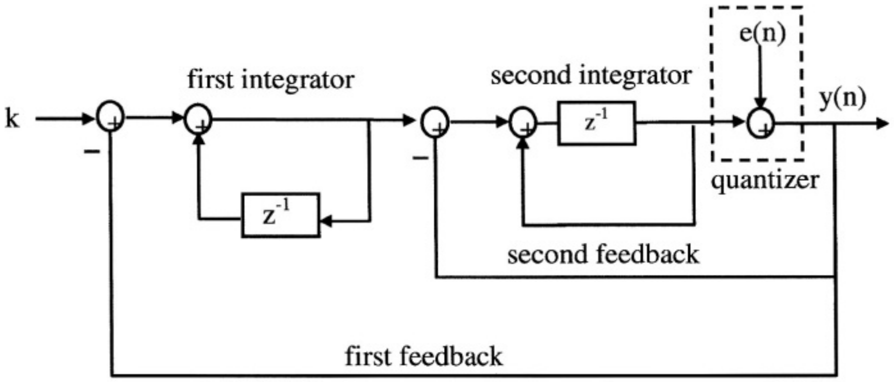


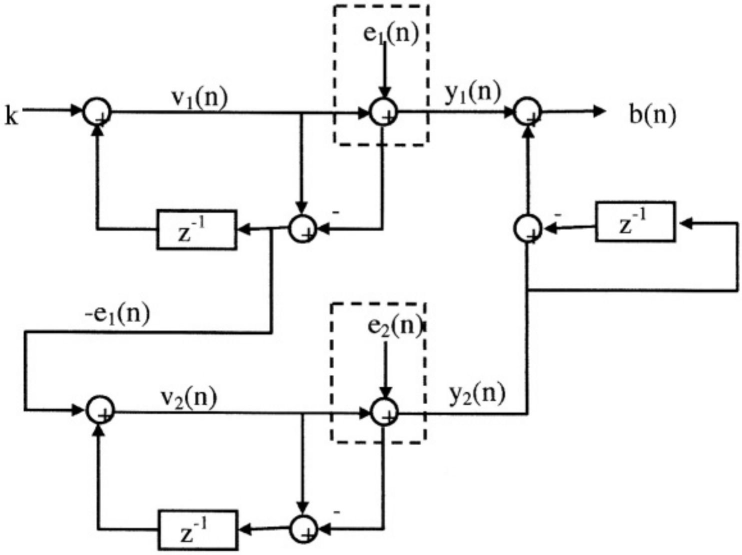
Fig. 2. Second order single loop ΔΣ modulator block diagram

The feedback type consists of digital integrators placed within a feedback loop of various complexity. The second order single loop modulator in Fig.2. is very tolerant to circuit imperfections when an inherently linear 1-bit quantizer is used. The first feedback determines the low frequency properties of the loop whereas the second feedback determines the high frequency properties and stabilizes the system. The noise at the output is shaped approximately as the opposite of the filter transfer function designed by the two digital integrators. The filter can have multiple poles at DC or poles distributed through the signal band when modulator order is greater than three. As such, there is high degree of flexibility to allow shaping of the quantization noise. The drawback of this structure is that third order modulator and above can become unstable due to reduced loop gain when quantizer is overloaded by large input signal levels. Actually, instability occurs when quantization error is amplified in the feedback loops and causes the quantization error to exceed $\pm\Delta/2$ where Δ is the quantization step.

Mainly, there are two techniques to avoid overload. The first one is to restrict the allowed input range below full scale. Consequently, the ΔΣ modulator order will increase to cope with dynamic range reduction and the penalty will be an increase of the die area. The second one requires the use of a multiple level quantizer. Usually, for a L^{th} order digital ΔΣ modulator, 2^L levels are required. This latter technique preserves the dynamic range with the restriction of a more complex quantizer. Most of the time the resulting L output bits do not accommodate for the needed range of the ΔΣ synthesizers when 2/3 dividers are used.

The second type of high order ΔΣ modulator is the cascade or MASH architecture. This topology uses combinations of inherently stable 1st and/or 2nd order ΔΣ modulators configured in a feedforward scheme to progressively increase the order of the whole modulator. Thus, the MASH topology is inherently stable. In addition, in ΔΣ synthesizers, the ΔΣ modulator is digital. Then, MASH architecture in digital circuit does not suffer from matching requirement as it would in an analogue implementation. The MASH output is multi-bit using single-bit quantizers blocks. So, with this structure it is easier to make fractional-N PLL with programmable fractional dividers. As illustrated by Fig.3, a MASH-1-1 that consists of two 1st order ΔΣ modulator cascaded will drive at least 2 programmable dividers.

This discussion suggests that in digital system the cascade architecture are favoured because they can achieve the same dynamic range at lower order than single-loop modulators and therefore are less power and silicon hungry.



a)

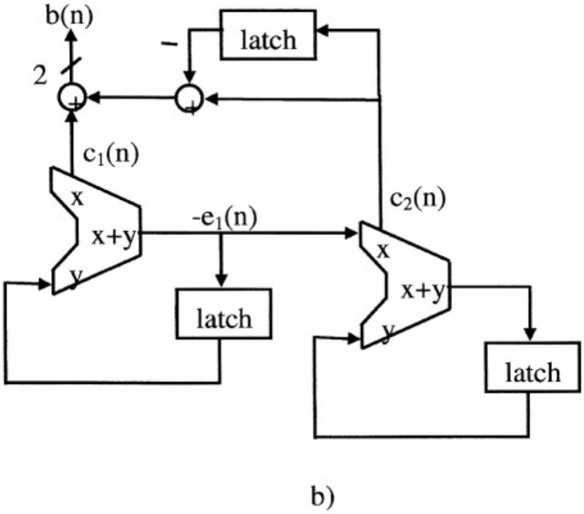


Fig. 3. Second order all digital MASH-1-1 modulator a) block diagram and b) its digital implementation.

3.2. ΔΣ modulator clock frequency

Over one period of the reference clock, the divider is unresponsive to the ΔΣ modulator output. Hence samples may be ignored if the ΔΣ clock was run faster than the reference. So an important constraint is that the frequency of the reference oscillator F_{REF} must be exactly equal to the ΔΣ clock frequency. It should be emphasised that the ΔΣ modulator only drives the divider ratio. The actual change takes place only when the divider has completely finished its preceding cycle count. In practice the divider must then run one sample behind the output of the ΔΣ modulator.

3.3. Reference frequency

With regard to the choice of reference frequency there are two competing trends that need to be discussed. The first is associated with the PLL. By doubling the reference frequency, the tuning range is also doubled and therefore the phase noise is increased by 6dB. The second trend is associated with the ΔΣ modulator. By doubling the clock frequency the noise density is decreased by $10.log\left(2^{2L-1}\right)$ where L is the modulator order. Hence the two trends compete with the net effect that phase noise on the VCO decreases by $10.log\left(2^{2L-3}\right)$ for every doubling of the clock or reference frequency. In the end, the higher the reference frequency, the better the noise performance but at the expense of more power consumption.

3.4. Digital accumulator size and topology

The digital accumulator size is mainly defined by the desired frequency resolution F_{STEP} defined by:

$$F_{STEP} = \frac{F_{REF}}{2^k} \tag{1}$$

where k is the number of bits required for the accumulator. Hence, high resolution requires a large number of bits. Adders can easily be realised using full adder logic circuit connected in cascade. This ripple carry adder configuration is a good choice for low noise injection in the substrate and power supplies. Actually, full adder circuitry is expected to change its output when carry of previous full adder stage is propagated. Then noise from gate switching is also cascaded or spread in time. However, carry propagation between each full adder may significantly reduce the operation speed of the overall design. Cascading carry look-ahead adders that anticipate carry propagation could be a solution. However, the hardware complexity exponentially increases with the number of bits at the output of carry look-ahead adders [10]. A simpler approach to break the long carry chain in ripple carry adders is to employ a pipelining technique [5] so that the carry information is only forwarded for example on half of the number of required full adders. Consequently, carry propagation time, is divided by two. A four-bit example of such a pipelined accumulator is shown in Fig.4. Since the ΔΣ modulator input is constant for the entire time, the input and output alignment registers, usually needed in pipelined structures, can be completely removed and substantial die area is saved. It has to be noted that when the modulator input changes dynamically as it does with ΔΣ fractional-N synthesizers used for in loop modulation then input and output alignment registers are needed for time synchronisation.

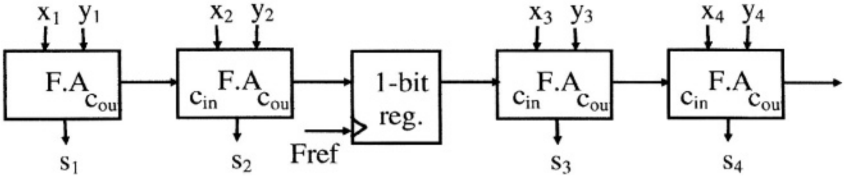


Fig. 4. Pipelined 4-bit adder formed out of two 2-bit ripple carry adders.

3.5. Designing tones free modulator

One of the drawbacks of ΔΣ modulators is the generation of tones in the output spectrum. These tones can lead to degraded phase noise performances at PLL output if there are not sufficiently removed.

When the modulator is fed with a DC-input, α, the modulator spectrum contains very strong out-of-band tones near $F_{REF}/2$ at frequencies [3], [6].

$$F_{TONES} = \left(1 - \frac{k|\alpha|}{\Delta} \right) \frac{F_{REF}}{2} \qquad k = \{1,2,3,\ldots\} \tag{2}$$

and small in-band tones at

$$F_{TONES} = \frac{F_{REF}}{2} k \frac{|\alpha|}{\Delta} \qquad k = \{1,2,3,\ldots\} \tag{3}$$

where Δ is the quantization step.

Increasing the order of the modulator will reduce in-band tones as well as in-band noise since this allows a higher in-band loop gain. Actually, when *K* fractional word is close to zero or one, then higher order single loop modulators are able to better decorrelate the successive output samples. However, whatever order the modulator is, the out-of band loop gain must remain nearly the same to preserve loop stability. Consequently the modulator order has no effect on the amplitude of tones near $F_{REF}/2$.

Another approach for tones suppression is the introduction of a degree of “uncertainty” into the modulation. This is accomplished either by adding a dither noise into the modulator loop or by the use of chaotic modulator [8], [9]

A widely used dithering technique is to add an extra least significant bit (LSB) always at one to the K fractional word. This systematic offset avoids repetitive cycle of carry generation in the accumulator and then tones. Unfortunately, in the presence of fractional word close to zero or one, when either ΔΣ modulator is a MASH one or a low order single loop one, tones appear at the modulator [7] outputs because a strong correlation exists between the output of the modulator and its input. In other words, the ΔΣ output is not a pseudo-random bitstream any longer. Then a pseudo-random dither signal is generally added to the input of the single-bit or multi-bit quantizer as illustrated on Fig.5. It can break up idle tones in the quantization noise floor and it is conveniently noise-shaped by the ΔΣ loop. In most of the cases, the dither is generated digitally using a maximum-length sequence generator also called PN sequence. However, the introduction of dither reduces ΔΣ modulator dynamic range. Indeed, the increase in the quantizer input noise power tends to decrease quantizer gain. Consequently, ΔΣ loop gain decreases and stability problems occur. In [8] it has been shown that the signal to noise ratio (SNR) penalty is 4 to 5 dB and by comparison with triangular probability density (TPD), rectangular probability density (RPD) dither achieves linearisation with marginally lower SNR costs. Then PN generator is often used with multi-level quantizers structures.

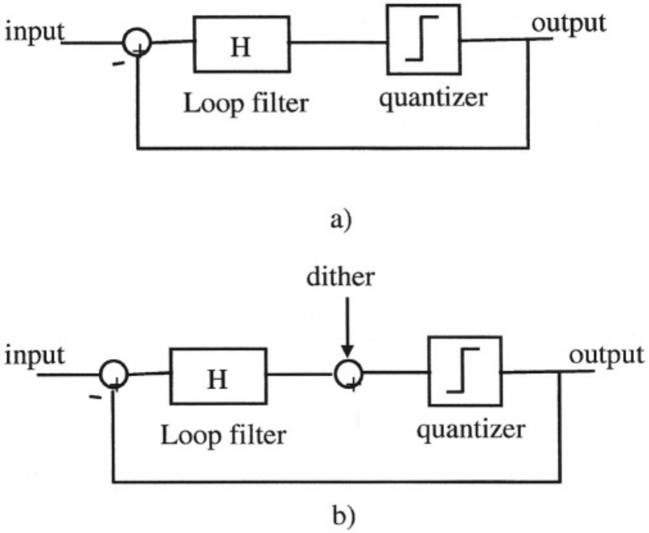


Fig. 5. General $\Delta\Sigma$ modulator: a) without, b) with dithering signal.

Chaotic ΔΣ modulator is an alternative to dither techniques to linearise ΔΣ loop [8][9]. The modulator is made chaotic when noise-shaping zeros are moved outside the unit circle in the z-domain which is equivalent to make H(z) open loop unstable. However, high level of chaos has been found necessary to eliminate tones and reduction of SNR can be up to 65dB for a 4th order modulator [8].

3.6. Phase-Frequency Detector (PFD) influence.

In a ΔΣ fractional-N synthesis, the linearity of the loop is mainly determined by PFD linearity. It has previously been observed that ΔΣ modulators suffer from tones. As tones near $F_{REF}/2$ are very energetic and are filtered after entering the PFD, two closely spaced tones at high frequencies can potentially result in their difference base-band tone because of intermodulation distortion due to non-linearities of the PFD. This folding mechanism leads to a rise in noise floor too because a part of huge quantization noise near $F_{REF}/2$ “leak” in the PLL bandwidth. This folding noise and tones are now low frequency and are not filtered any longer by the loop filter. Consequently, the close-in phase noise of the PLL is increased and strong tones near carrier frequency appear. To avoid such a problem, a delay is added to the PFD. Therefore, both the up and down charge pump are 'on' for a short period of time when near-zero or zero phase error is detected. This technique enables zero-dead zone suppression and linearises the PFD. However, any mismatches between the up and down current will result in a small charge injection in the loop filter at each cycle and will generate reference sidebands tones at VCO output that will tend to increase the amplitude of the ΔΣ tones. In presence of dithering, mismatches might again generate spurious and increase the close to carrier noise floor [11]. Typically, better than 5% matching for narrow-band modulation formats is needed to maintain close in phase noise at a low level.

4. Architecture using ΔΣ fractional-N PLL.

4.1. Frequency synthesis

4.1.1. Noise/lock time

As said above, fractional-N PLLs are more and more used in modern RF transceivers due to their superior performance in terms of noise and or lock time. This is especially true for a ΔΣ fractional-N PLL because noise-shaping behaviour allows to achieve near ideal noise performance. As almost any frequency step can be realised through the fractional value, the comparison frequency can be maintained relatively high which improves both the close in phase and lock time. One can widen the loop bandwidth for a given phase error or jitter compared with integer N PLL (shorter lock time) or for the same loop bandwidth improve phase error or jitter (better close in phase noise). This has to be tempered by the fact that more noise is generated at higher frequency offsets that requires very careful loop filter design.

4.1.2. Automatic Frequency Control through a fractional-N PLL

Cellular networks requires near perfect frequency alignment (down to 0.1 ppm) between the mobile and the base station, this is traditionally achieved in the mobile by frequency tuning of the reference crystal oscillator. This requires a varactor diode and a specific DAC in the system. This extra hardware can be avoided by the use of ΔΣ fractional-N PLLs since the frequency step F_{STEP} can be made as small as required to align the RF VCO frequency to the base station (equation 1). As an example a 17 bits MASH accumulators is required to generate a 2GHz frequency with better than 0.1ppm accuracy from a 26 MHz frequency.

4.1.3 Multi-mode system clock

Another example of the use of fractional-N PLL technique is the generation of different system clocks for multi-mode application. For example GSM symbol rate (270kHz) and WCDMA chip rate (3.84MHz) cannot be derived from a single crystal oscillator by integer division. Rather than using two crystals (for example 13MHz and 19.2MHz) it is fairly easy and cheaper to generate 32.72MHz (8x3.84MHz) from a 13 MHz crystal with a fractional-N PLL. On top of this a separate AFC control for the two system clocks can also be implemented.

4.2. Frequency conversion and LO generation techniques enabled by fractional-N PLL

4.2.1. Variable IF offset modulation loop.

In the latest years the transmit side of the GSM phones have converged to an optimised low noise RF architecture avoiding use of lossy filtering after the power amplifier: the offset modulation loop. This technique applies the modulation directly to a low noise VCO in a phase locked loop where the reference frequency is a GMSK modulated IF frequency. Although being quite optimized this technique traditionally requires an IF frequency to be generated. Thus an extra VCO and PLL must be integrated in the system which is no longer needed for the receive side in modern transceiver using direct conversion. Fractional-N PLL allows to solve that problem with the variable IF offset loop architecture. This is a more robust alternative to the direct modulation via the fractional-N PLL but requires more hardware.

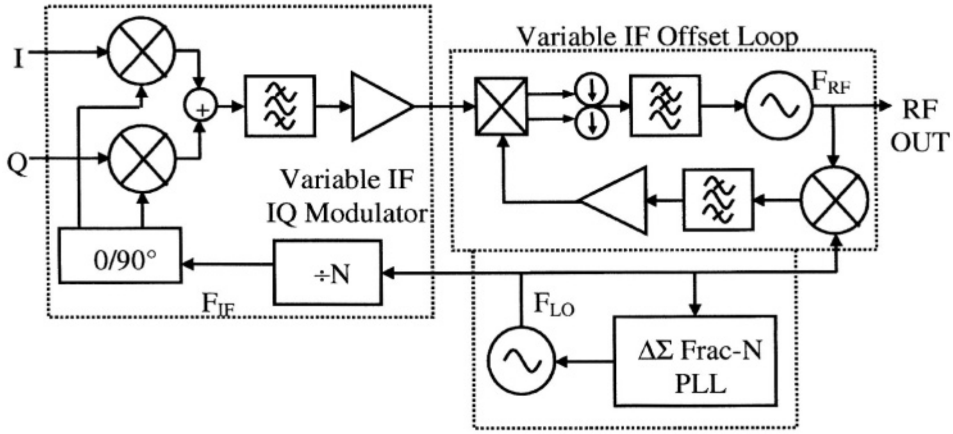


Fig .6: Variable IF offset modulation loop using ΔΣ fractional-N PLL

The principle is to use a frequency divided version of the LO to create a variable IF frequency as can be seen in Fig. 6. In this case, the output frequency step is a fractional value of the reference frequency of the PLL controlling the LO VCO as shown in equation 4, the sign in the equation depending on the side band selected in the down mixer.

$$F_{RF} = F_{LO} \cdot \frac{N \pm 1}{N} \tag{4}$$

In this case only fractional-N PLL allows synthesizing the correct transmit frequency.

4.2.2. Variable IF pseudo direct conversion.

A similar technique can be used in systems where the transmitter is based on super heterodyne architecture and the receiver is a direct conversion one.

The main reason to use such an architecture in transmit is for cases where the modulation contains both amplitude and phase information and need a dynamic range AGC for power control which must be splitted between IF and RF as in CDMA systems. Such architecture is illustrated in Fig.7. Here again the output frequency step is a fractional value of the reference frequency of the PLL controlling the LO VCO, the same equation 4 applies making it only possible if the VCO is controlled within fractional-N PLL. It is noticed that such a variable IF architecture can be used in receivers too.

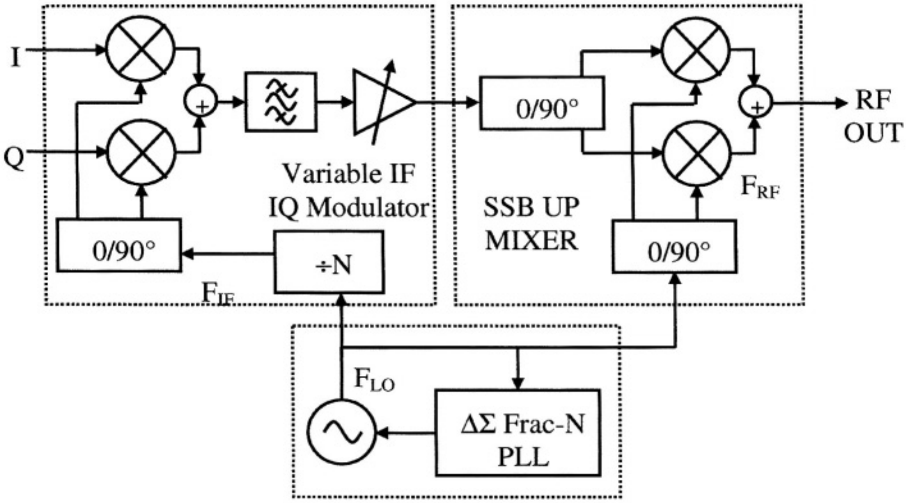


Fig.7: Variable IF Up-Mixer Transmitter using ΔΣfractional-N PLL

4.2.3. LO regeneration

In the future the cellular phones will have to cover more and more bands to offer international and multi-standard coverage to the user. In such a phone traditional techniques requires multiple VCOs to cover the different receive and transmit bands. This is quite inefficient and expensive. Cheaper techniques trying to extend a single VCO range to cover all bands are much more efficient. Two of these techniques known as feed-forward and feedback fractional LO generation are illustrated in Fig. 8. Here again the best architectures make use of fractional-N PLL to be able to synthesize the desired frequency with the best performance as equations 5 and 6 shows for the feed-forward and feedback LO generation respectively.

$$F_{LO} = F_{VCO} \cdot \frac{N \pm 1}{N} \tag{5}$$

$$F_{LO} = F_{VCO} \cdot \frac{N}{N \pm 1} \tag{6}$$

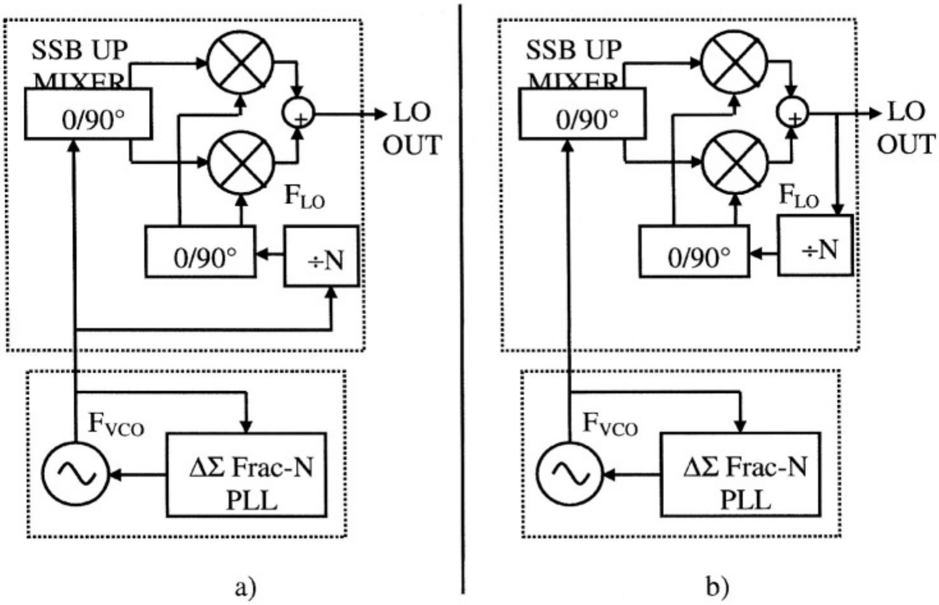


Fig.8: Fractional LO regeneration a) Feed-Forward type; b) Feed-Back type

Conclusion

In this paper, we discuss ΔΣ fractional-N PLL architecture trade-offs and implementations in highly integrated multi-mode circuits. The phase-frequency detector (PFD) non-linearities and charge pump (CP) mismatches are the main sources of close in phase noise degradation. Typically better than 5% matching charge pumps associated with zero-dead zone PFD are used. A MASH ΔΣ architecture built from 1st and/or 2nd order stages is always stable and is very attractive for digital implementations when pipelining technique is used, ΔΣ tones can be whitened thanks to digital dithering signal. Then, the complexity introduced by ΔΣ modulator is digital and is highly repeatable and robust. Implementation of ΔΣ fractional-N PLL in multi-mode (GSM-WCDMA) circuits enable to implement automatic frequency control on multi-mode system clock derived from a single low cost 13MHz crystal. In addition, variable IF in transmitter as well as LO regeneration needed for future multi-bands transceivers require a fractional-N PLL and lead to the reduction of the number of VCOs. Thus, systematic used of ΔΣ fractional-N PLL is very efficient in terms of die area and cost and will extensively be used in next generation of multi-mode transceivers.

References

[1] D. Brunel et al, "A Highly integrated 0.25um BiCMOS Chipset for 3G UMTS/WCDMA Handset RF Sub-System", IEEE RFIC symposium, 2002.

[2] B. Goldberg, *"Digital Techniques in Frequency Synthesis."* McGraw-Hill, 1995.

[3] S. R. Norsworthy, R. Schreier, G. C. Temes, *"Delta-Sigma Data Converters – Theory, Design, and Simulation",* IEEE Press, 1997

[4] T. Riley, M. Copeland and T. Kwansniewski, *"Delta-sigma modulation in fractional-N frequency synthesis,"* IEEE J. Solid-State Circuits, ol. 28, pp.553-559, May 1993.

[5] M. H. Perrot, "Techniques for high data rate modulation and low power operation of fractional-N frequency synthesizer with noise shaping", Ph.D. thesis, M.I.T, Cambridge, MA, 1997

[6] L. Breems, J.H. Huijsing, *"Continuous Time SD modulation for AD conversion in Radio Receivers"* Kluwer Academic Publisher, 2000, ISBN 0-7923-8625-4

[7] B. De Muer, M. S. J. Steyeart *"A CMOS Monolithic DS-Controlled Fractional-N Frequency Synthesizer for DCS-1800"* IEEE J. Solid-State, vol. 37 pp 835-844, July 2002.

[8] C. Dunn and M. Sandler *"Linearising Sigma-Delta Modulators using Dither and Chaos"* proc. ISCAS, 1995 pp 625-628.

[9] L. Risbo, *" Σ–Δ Modulators - Stability Analysis and Optimization",* Ph. D. Thesis, Technical University of Denmark, Chap.8, pp 123-132, 1994.

[10] M. Kozack, I. Kale *"A Pipelined Noise Shaping Coder for Fractional-N Frequency Synthesis",* IEEE Trans. Instrum. Meas., vol. 50, pp1154-1161, Oct. 2001.

[11] D. Barclay, *"Fractional-N PLLs – Theory, Application and Practices",* Proc. RF Circ. Techn. workshop, Cambridge, March 2002

FRACTIONAL-N PHASE LOCKED LOOPS AND IT'S APPLICATION IN THE GSM SYSTEM

Günter Märzinger, Burkhard Neurauter
Danube Integrated Circuit Engineering
Freistädter Straße 401
A-4040 Linz, Austria

Abstract

This paper presents an overview of the fractional-N frequency synthesis technique. After a short excursion into the history of ΔΣ modulation, different fractional-N techniques are shown. Following this, the ΔΣ fractional-N PLL is explored in detail, and important results relating to fractional-N PLLs are presented. In the next section the blocks of the analogue part of the PLL are listed, and relevant design criterions are discussed. Finally the specifications imposed by the GSM system are mentioned, and two techniques for using a fractional-N PLL as a GMSK modulator are shown.

1. Introduction

In the last years the mobile communications industry has experienced a change into a consumer market. 2G and 2.5G systems are dominating the scene, while 3G is still far out at the horizon. Therefore low cost, high efficient and highly integrated transceiver architectures are being investigated by the industry. In recent years the technique of fractional-N frequency synthesis caught strong attention by mobile phone chip makers, because it offers the potential for highly integrated and current saving transmitter architectures. But still the fractional-N PLL is not a common found ingredient in communication ICs, like it is the case with integer-N PLLs. This might be because in designing good fractional-N PLLs three areas of expertise have to be covered. Namely digital signal processing, control theory and analogue circuit design.

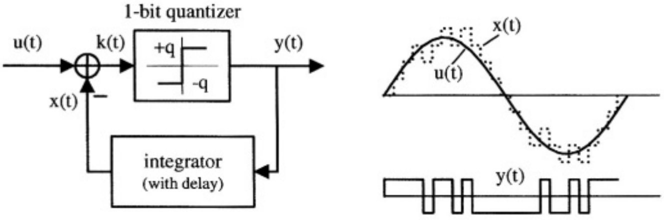
A. van Roermund et al. (eds.), Analog Circuit Design, 111-127.
© 2003 *Kluwer Academic Publishers. Printed in the Netherlands.*

2. ΔΣ modulation

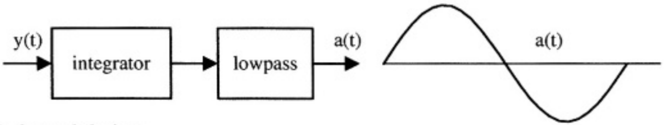
The basic principle of ΔΣ modulation is the usage of feedback and over sampling to improve the resolution of a coarse quantizer.

2.1. History of ΔΣ modulation

Already in the year 1952 de Jager presented the delta modulator [1]. At this time he intended to use this principle for communication purposes. Figure 1 shows this communication technique.


1-bit quantizer
u(t)
k(t)
+q
-q
y(t)
x(t)
−
integrator
(with delay)
x(t)
u(t)
y(t)

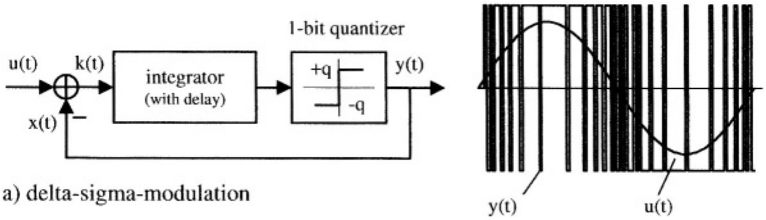
a) delta-modulation


y(t)
integrator
lowpass
a(t)
a(t)

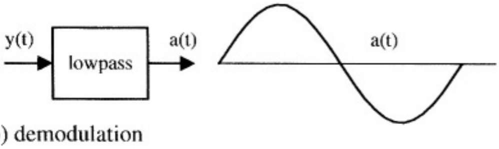
b) demodulation

Fig. 1. a) Delta modulation, b) demodulation.

The sent data stream y(t) was 'modulated' in a manner, that only changes in the input signal were transmitted, hence delta modulator. At the receiver a simple integrator and a low pass filter could reconstruct the source signal u(t).


1-bit quantizer
u(t)
k(t)
integrator
(with delay)
+q
-q
y(t)
x(t)
−
y(t)
u(t)

a) delta-sigma-modulation


y(t)
lowpass
a(t)
a(t)

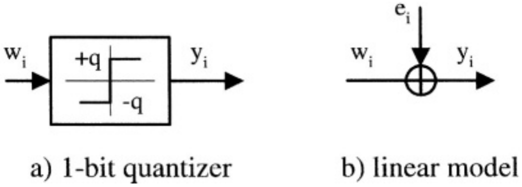
b) demodulation

Fig. 2. a) Delta-Sigma modulation, b) demodulation.

As the integration is, in a mathematical sense, a linear operation, it was suggested to put the integrator of the demodulator before the delta modulator and then shift it into the loop [2]. The result is depicted in figure 2. As this system was constructed from a delta modulator and an integrator it was called delta-sigma-modulator. Nowadays the notation ΔΣ and ΣΔ modulation are both in use.

Although it is possible to get an intuitive feel for the working principle of ΔΣ modulators from their behavior in the time domain, the easiest way to understand a ΔΣ modulator is in terms of the frequency domain description of its linear model. In this linear model, the nonlinear operation of the quantizer is replaced by a (linear) addition of a random error sequence. Linear system theory can than be utilized to show that the output of the modulator is the sum of the (filtered) input signal and the filtered quantization noise. The two components can be filtered independently under the control of the designer. This allows ΔΣ modulators to separate spectrally the input signal from the quantization noise.

Although figure 2 is a continuous time system, the following math is done in discrete time. The key to the linear model is the substitution of the nonlinear quantizer by a linear addition of an error signal e_i as shown in figure 3.



a) 1-bit quantizer b) linear model

Fig. 3. Linear model of nonlinear quantizer.

The error signal e_i is entirely determined by the input signal w_i. But under certain conditions [3] e_i can be modeled as uncorrelated random signal with uniform distribution within the quantization levels

$$q = \pm\frac{\Delta}{2}. \tag{1}$$

Therefore the power of the error signal is given by

$$\overline{e^2} = \frac{1}{\Delta}\int_{-\frac{\Delta}{2}}^{+\frac{\Delta}{2}} e^2 de = \frac{\Delta^2}{12}, \tag{2}$$

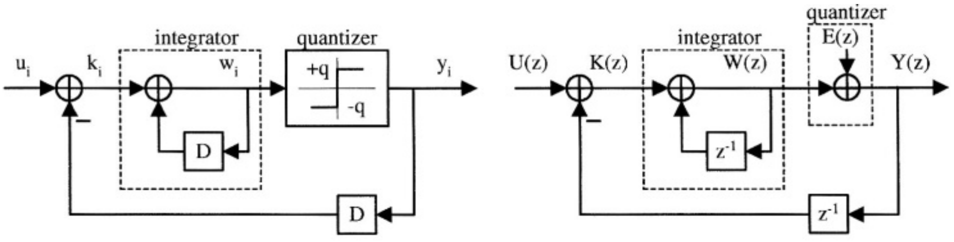
and the power spectral density can be calculated as

$$S_{ee(f)} = \frac{\overline{e^2}}{f_s} = \frac{\Delta^2}{12 f_s} \tag{3}$$

with f_s as the sampling frequency.

2.2. First order delta sigma modulator

Figure 4 a) shows a discrete time realization of a first order ΔΣ modulator, and figure 4 b) the associated linear model in the frequency domain.



a) discrete time realization

b) linear model

Fig. 4. Discrete time realization and linear model of a 1^{st} order ΔΣ modulator.

In the time domain the difference equation of the modulator is given by

$$y_i = w_i + e_i = u_i + e_i - e_{i-1} \tag{4}$$

which in the frequency domain results to

$$Y(z) = U(z) + E(z) \cdot (1 - z^{-1})\,. \tag{5}$$

Equation (5) shows clearly the spectral shaping of the quantization noise by a first order high pass. It is also possible to design higher order noise shaping. A more detailed insight into theory can be found in [4] and [5]. Especially if one designs higher order modulators the problem of stability arises.

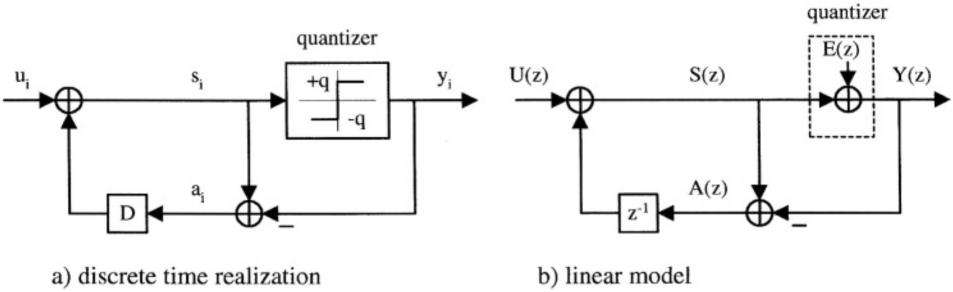
2.3. The Error feedback structure

Up to now a variety of different ΔΣ modulator structures were designed. In fractional-n frequency synthesis the error feedback structure shown in figure 5 has gained some importance.

As a_i=-e_i the difference equation of the error feedback structure in the lime domain is given by

$$y_i = s_i + e_i = u_i + e_i - e_{i-1}\,, \tag{6}$$

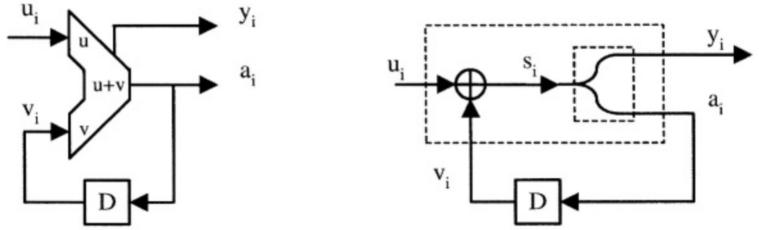
which is the difference equation of a first order ΔΣ modulator.



a) discrete time realization b) linear model

Fig. 5. Discrete time realization and linear model of an error feedback structure.

It is now an interesting fact, that a simple digital accumulator is indeed the same as the above mentioned error feedback structure. If the sum of the accumulator, as depicted in figure 6, is split up in a MSB part (y_i) and a LSB part (a_i), and only the MSB part (in the extreme only the carry) is used for further processing, than this is equal to a quantization. Thus a_i is again equal to the negative quantization error and therefore the equivalence of a digital accumulator and an error feedback structure is given.



a) digital accumulator b) quantization of the ideal sum by truncation of the lower bits

Fig. 6 Digital accumulator as error feedback structure.

3. Review of the basic LTI model of PLLs

In order to have the mathematical tooling to be able to predict important PLL parameters, commonly a linear, time invariant model of the PLL and its building blocks is used [6], [7]. As it is a quite common practice to design synthesizers having a current output at their phase detector, and utilizing an oscillator needing voltage control, the loop filter must act as trans impedance. Adding internal noise sources to all the blocks, the model as depicted in figure 7 results [8], [9].

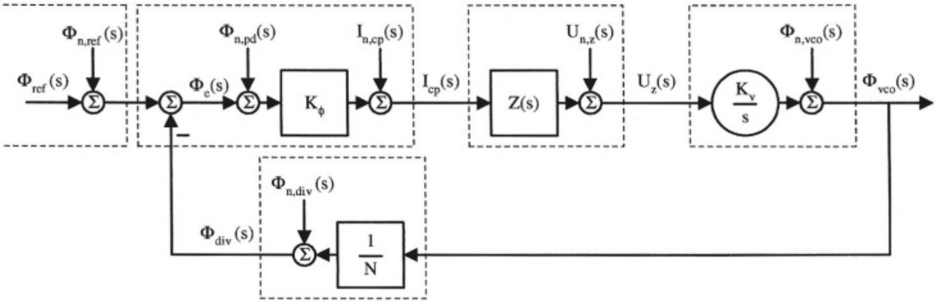


Fig. 7. LTI model of a PLL frequency synthesizer with additive noise sources.

Using this model the open loop transfer function can be defined as

$$L(s)=\frac{\Phi_{div}(s)}{\Phi_{ref}(s)}=\frac{K_\Phi K_v}{N}\frac{Z(s)}{s} \tag{7}$$

and the closed loop transfer function is given hy

$$G(s)=\frac{L(s)}{1+L(s)} \tag{8}$$

and therefore

$$\frac{\Phi_{vco}(s)}{\Phi_{ref}(s)}=N\cdot G(s)\,. \tag{9}$$

4. Fractional-N PLLs

Indirect frequency synthesis with the aid of PLLs is already a well established technique. But up to the present mainly integer-N designs were used. As can be seen from (9) the VCO phase is the reference phase multiplied by N. Therefore also the VCO frequency is the reference frequency multiplied by N. This has two major consequences. First only integer multiples of the reference frequency can be controlled in a single loop [6], [7]. Thus the name integer-N PLL. Second also the phase noise of the reference frequency as well as the phase noise of the divider and the phase detector are multiplied by N. If a system using integer-N PLLs needs to have a fine frequency resolution the signal of the controlled VCO becomes noisy, as N becomes very big. Therefore, if the quality of the PLL building blocks can't be increased anymore, the only way to get a better PLL output signal, is to reduce the bandwidth of the control loop G(s).

4.1. Different fractional-N PLL techniques

As the above mentioned trade off between speed of the control loop, which is direct proportional to the bandwidth of G(s), and noise performance, which is in a system with a given reference frequency indirect proportional to the bandwidth of G(s), always has given the engineers a hard time, the wish to decouple the frequency resolution of the synthesizer from the phase detector comparison frequency exists already for a long time.

As a frequency divider in the feedback path of the control loop can only divide by integer numbers, the basic idea of fractional-N division is to obtain a fractional number as a time average out of a sequence of integer numbers. The task of time averaging is performed by the limited speed of the PLL. This gives a first intuitive demand for fractional-N PLL design, namely that the phase detector comparison frequency has to be high in relation to the bandwidth of the control loop. It is easy to see, that a divider which is dividing a signal having a constant input frequency by a sequence of integer numbers, will produce an output signal having the wanted frequency on an average, but this output signal will show a great portion of jitter. This jitter is passed through the phase detector, and filtered out to some degree in the loop filter. In [10] fractional-N configurations are classified into first, second and third order FN-PLLs, according to the way they deal with jitter generation/cancellation. To avoid confusion with the term 'order' in this paper fractional-N techniques should be classified with first, second and third 'kind'.

4.1.1. Fractional-N of the first kind

The first step towards a fractional-N synthesizer is to find an appropriate algorithm for producing a sequence of divider values having the wanted fractional number as a time average. The simplest way to do this is by using the carry of an accumulator for selecting the wanted divider value. (Remark: Generally the divider value could be changed between N_0 and $N_0+\Delta N$. But for the sake of simplicity, without loss of generality, ΔN is chosen to be 1).

Using this technique a fractional-N PLL as depicted in figure 8 results. This synthesizer is classified as fractional-N of the first kind.

Mathematically the operation of a digital accumulator with n bit width can be expressed as an addition modulo 2^n. If K is the number applied to the input of the accumulator, and $F=2^n$, then

$$v_{i+1} = (v_i + K) \bmod F \, . \tag{10}$$

It is a property of the modulo operation, that

$$\forall x \in Z: \quad (x \cdot F) \bmod F = 0 . \tag{11}$$

Because $$(a+b) \bmod F = (a \bmod F + b \bmod F) \bmod F \tag{12}$$

and $F \cdot K = K \cdot F$ the longest cycle the accumulator can go into is F clock cycles. $$v_{i+F} = (v_i + F \cdot K) \bmod F = v_i . \tag{13}$$

And because $$(v_i + F) \bmod F = v_i \tag{14}$$

means that the accumulator has reached its initial state and has produced one carry bit, then for any K the accumulator produces K carry bits in F clock cycles and ends up in its initial state. Therefore the cumulated divider value is

$$N_{cum} = K(N_0 + 1) + (F - K)N_0 = K + F \cdot N_0 , \tag{15}$$

which gives an average divider value of

$$\overline{N} = \frac{N_{cum}}{F} = N_0 + \frac{K}{F} . \tag{16}$$

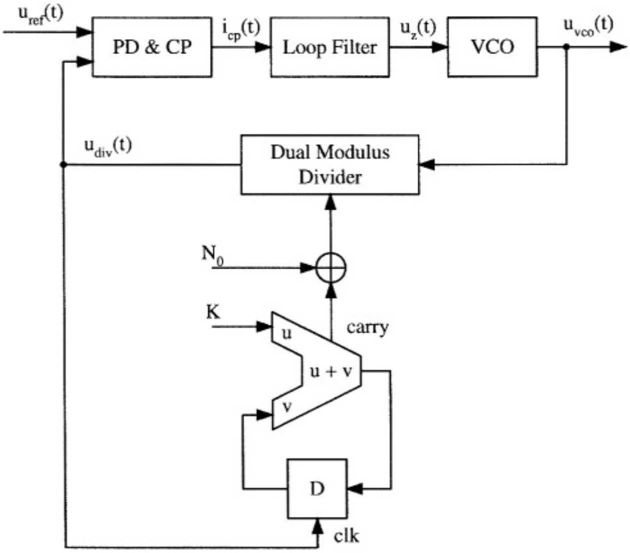


Fig. 8. Fractional-N synthesizer of the first kind.

Because the sequence of divider values is cyclic with a longest period of F clock cycles (but it can be shorter), the phase jitter at the divider output shows a great spurious content. These so called fractional spurious are only filtered by the loop filter and are a great disadvantage of fractional-N synthesizers of the first kind. Therefore this technique is not very popular. An example is [11]. In [12] this method is only used to obtain a fast lock time, in the locked condition this PLL is switched back to integer mode.

4.1.2. Fractional-N of the second kind

As the spurious performance of a first kind fractional-N is not sufficient for most applications efforts have been made, to cancel out the jitter introduced by the switching of the divider. Since the accumulator content is a measure of the instantaneous phase error, this information can be used to perform a cancellation [6], [13]. Feeding the accumulator content into a digital to analogue converter (DAC) to convert it to a proper scaled analogue correction value, this signal can, in theory, cancel out the phase jitter, by superimposing it to the charge pump output.

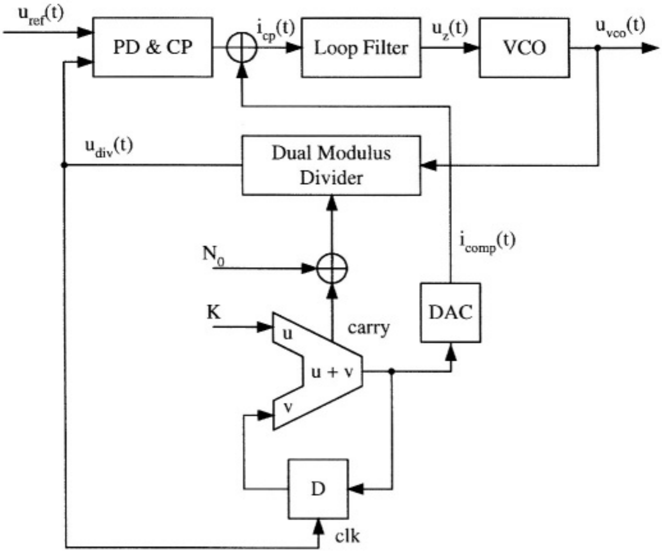


Fig. 9. Fractional-N synthesizer of the second kind.

Using this technique a fractional-N PLL as depicted in figure 9 results. This PLL is classified as fractional-N of the second kind.

In practical applications the precision of this analogue cancellation scheme is limited, and it is very sensitive to process tolerances, aging and temperature drift. Though fractional-N of the second kind found more applications in industry [14], it could not revolutionize the method of indirect frequency synthesis. It should also be mentioned, that the idea of using a phase modulator at the dual modulus divider output was investigated too, in order to cancel out the phase jitter [15].

4.1.3. Fractional-N of the third kind

In the chapter about ΔΣ modulation it was shown, that an accumulator is equivalent to a first order ΔΣ modulator. Therefore the problem of avoiding fractional spurs is the same, as the problem of avoiding idle tones in ΔΣ modulation. Furthermore all the theory and the results obtained in the vast field of ΔΣ digital to analogue conversion can now be applied to frequency syntheses. Because in a special sense, a PLL is a kind of digital to analogue converter. Thinking about N as the digital number applied to the 'converter', and the frequency of the VCO as the analogue output signal, this becomes obvious. With this in mind, an all digital solution to high performance fractional-N frequency synthesis can be found.

Already in [16] and [17] this all digital fractional-N technique was published. They used a 3rd order MASH modulator for generating the divider sequence. Using this technique a fractional N-PLL as depicted in figure 10 results. This PLL is classified as fractional-N of the third kind.

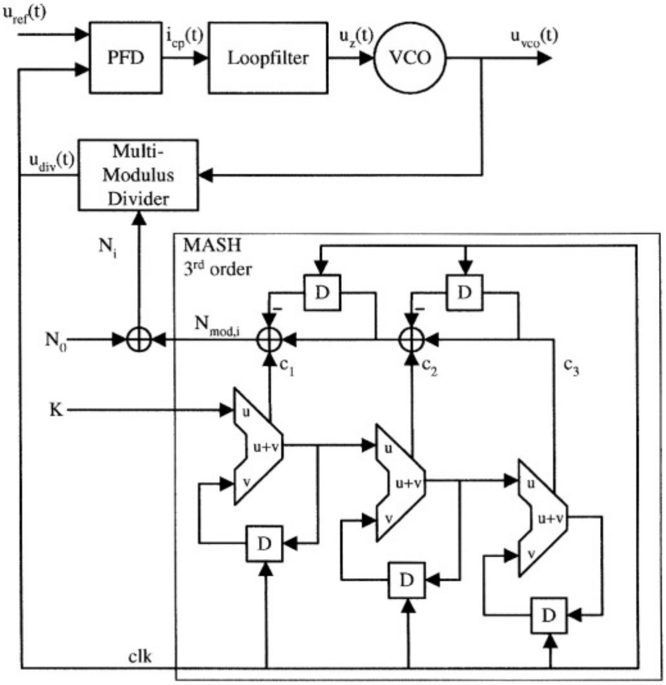


Fig. 10. Fractional-N synthesizer of the third kind.

As mentioned above, higher order modulators show the problem of stability. In [18] a technique called MASH (multi stage noise shaping) was presented. This technique is based on cascading stable first order modulators, and using the quantization error of the previous stage as the input signal of the succeeding to obtain an inherent stable modulator of higher order. The outputs of the different stages are merged in a manner that the quantization errors of the first stages cancel out, and only the quantization error of the last stage remains, filtered with a high pass function of order n, where n is the number of stages cascaded. It is easy to see, that

$$C_1(z) = \frac{K}{F} + E_1(z)\cdot(1-z^{-1})\,, \tag{17}$$

$$C_2(z) = -E_1(z) + E_2(z)\cdot(1-z^{-1})\,, \tag{18}$$

$$C_3(z) = -E_2(z) + E_3(z)\cdot(1-z^{-1})\,, \tag{19}$$

$$N_{\mathrm{mod}}(z) = \frac{K}{F} + E_3(z)\cdot(1-z^{-1})^3\,. \tag{20}$$

Now the output frequency of the PLL can be calculated by

$$f_{vco}(z) = f_{ref}\left(N_0 + \frac{K}{F}\right) + f_{ref}E_3(z)\cdot(1-z^{-1})^3\,. \tag{21}$$

Equation (21) shows, that the output of the PLL is composed of the reference signal multiplied by the wanted fractional number plus the quantization noise shaped by a third order high pass. As in synthesizer design the phase fluctuation, and not the frequency fluctuation of the output signal is a measure of quality, the noise term has to be translated to phase. With the result from equation (21) and a few lines of math as shown in [17] the phase fluctuations introduced into the loop by the ΔΣ modulator are given by

$$S_{\phi\phi\Delta\Sigma,in}(f) = \frac{(2\pi)^2}{12 f_{ref}}\left[2\sin\left(\pi\frac{f}{f_{ref}}\right)\right]^{2(n-1)}\,. \tag{22}$$

This phase noise is filtered by the low pass transfer function of the PLL and appears at the output by

$$S_{\phi\phi\Delta\Sigma,out}(f) = S_{\phi\phi\Delta\Sigma,in}(f)\cdot\left|G(j2\pi f)\right|^2\,. \tag{23}$$

5. Important properties of ΔΣ modulators used in frequency synthesis

A) Obviously the ΔΣ modulator has to be stable.

B) A certain frequency range has to be covered without gap.

C) The modulator architecture should not be susceptible to idle tones.

Using a MASH modulator requirement A) is no problem.

Several ΔΣ modulator architectures show a limited input range in order to have the wanted noise shaping, which is often restricted to lie between 0.2 and 0.8 [4]. Simulations confirmed, that a MASH modulator (with a sufficient number of bits in the accumulators; n>15) reaches the wanted noise shaping even with only the LSB set to one. Therefore a fractional-N synthesizer utilizing a MASH modulator can cover a frequency range in a simple way, even when it contains integer multiples of the reference frequency.

Equation 22 relies on the assumption, that the sequence of quantization errors is uncorrelated and uniformly distributed within the quantization interval. Simulations showed that for higher order MASH modulators this assumption holds to a high degree, but not for all K. Unfortunately till date, there is no theory that addresses the exact spectra of higher order ΔΣ modulators. But with a few lines of algebra it can be shown, that the cyclic sequence of the first accumulator in a MASH structure is the maximum of F for any constant K where the LSB set to one [19]. Under this boundary condition and using a MASH with about 20 bits in the accumulators the authors of this article encountered no fractional spurs, neither in simulation nor in measurement.

Beside these useful properties, the MASH modulator is easy to implement, even when pipelining is applied to reduce power consumption, and therefore it is the favorite choice in many designs. But there are also other modulator architectures in use [20], mainly because the MASH needs a quite wide modulo range in the N divider.

6. Important properties of PLL building blocks

It is important for equation 22 to hold, that the sequence of divider values, calculated by the ΔΣ modulator, is perfectly translated into phase jitter by use of the N-divider. If the delay of the N-divider is not the same for all the divider values used by the N-sequence, then the property of noise shaping is lost to some degree, and the noise close to the carrier increases.

Furthermore the phase jitter at the phase detector input needs to be translated exactly into current pulses by the phase detector – charge pump block. At least three mechanisms can deteriorate the phase-to-current translation. First the phase detector is not allowed to have a dead zone, as obviously this is a nonlinearity. Second, if a conventional loop design with an integrating loop filter is used, the up and down currents need to match to a high degree in their amplitude, because otherwise a nonlinearity due to the buckling at the origin would arise. And third, the charge pumps are not allowed to have any dynamic characteristic when they are turned on or off respectively, because this can also cause a non linear relation between phase and charge. Therefore, if a conventional phase-frequency detector together with an integrating loop filter is used, the operating point of the phase detector needs to be shifted from the origin by means of an offset current source [19]. Alternatively, another phase detector – charge pump topology can be used [21].

The last PLL block, which needs some attention, is the loop filter. If it is implemented by passive components, the linearity is no problem. But if it is implemented by some active circuitry, the needed linearity can become an issue.

7. The beneficial use of fractional-N PLLs in GSM ICs

Having a frequency synthesizer with arbitrary fine frequency resolution, this suggests itself to a minimum frequency modulator architecture [22]. As GSM uses a GMSK modulation scheme [23] with 270.83 kbit/sec, a modulator as shown in figure 11 would, in principle, be possible.

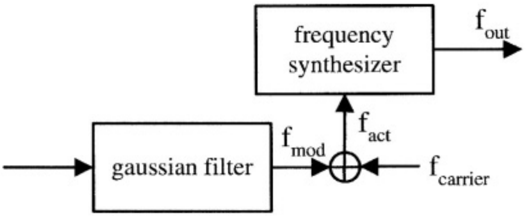


Fig. 11. Minimum modulator architecture for a GSM transmitter.

When designing a modulation loop to be used in a GSM mobile phone, phase noise requirements for 400 kHz offset, which is -113 dBc/Hz, and for 20 MHz offset, which is -162 dBc/Hz, are of special importance [24]. These two requirements impose restrictions to the usable PLL bandwidth. First, the bandwidth of G(s) has to be narrow enough, that equation 23 results to a value well below -162 dBc/Hz at 20 MHz offset. This is because in the noise budget of a GSM transmitter IC these -162 dBc/Hz are fully consumed by the VCO and buffer blocks. Second, as the achievable in band noise performance of a fractional-N design using 13 or 26 MHz as a reference frequency, is somewhere in the region of -90 to -100 dBc/Hz [25], the bandwidth of G(s) again has to be narrow enough, in order to fulfill the -113 dBc/Hz at 400 kHz offset. Therefore a reasonable PLL bandwidth is in the range from 80 kHz to 100 kHz. As all the signals fed into the PLL via the N-divider are filtered by the low pass transfer function G(s), also the modulation signal is low pass filtered. Figure 12 shows the resulting RMS phase error as a function of the -3 dB bandwidth of the PLL, when using the conventional passive 5 element integrating loop filter. It is apparent, that fulfilling the noise requirements will result in a PLL which is to narrow for a GSM signal (**$\text{phi}_{\text{rms,max}}$=5°** [23]). Thus measures have to be taken, to overcome this restriction.

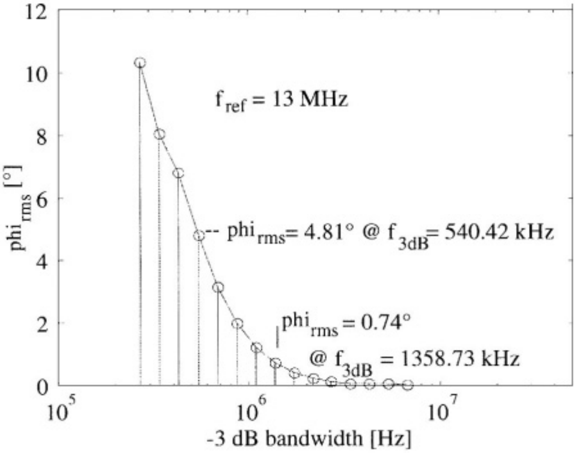


Fig. 12. RMS phase error versus -3 dB bandwidth of G(s).

In [21] a predistortion of the modulation signal is suggested, in order to overcome the bandwidth limitation which is given by the PLL design. The problem with this method is, that the predistortion and the PLL transfer function have to match to a high degree. As the PLL is an analogue control system, its parameters may vary in production, with temperature and aging. The poles and zeros can be controlled to have reasonable tolerances, by either using precise external R's and C's, or by using a switched capacitor loop filter as suggested in [21]. Therefore the worst analogue parameter is the open loop gain. Figure 13 shows the simulation result of the RMS phase error as a function of open loop gain error, when using the same PLL as mentioned above with about 80 kHz of bandwidth.

In [26] a two point modulation technique is presented. Two point modulation means, that not only the low pass path via the MASH modulator and the N-divider is used for modulation purposes, but also the signal is added after the loop filter. For this signal path a high pass transfer function according to

$$\frac{f_{vco}(s)}{U_{\mathrm{mod},z}(s)} = K_v \frac{1}{1+L(s)} = K_v\bigl(1-G(s)\bigr) \qquad (24)$$

results. In the 2-point modulator only the high pass gain

$$K_{HP} = K_{DAC}K_v, \qquad (25)$$

which is the conversion gain from the digital word to the analog frequency, has to be exact. Figure 14 shows the simulation result of the RMS phase error as a function of the high pass gain error, when using the same PLL as has been used to produce figure 13.

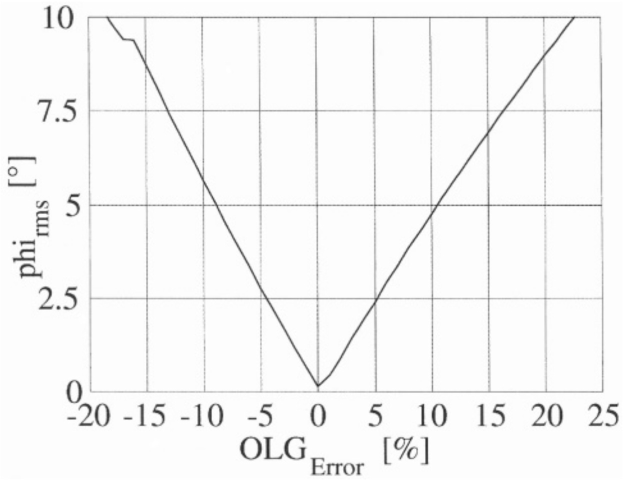


Fig. 13. RMS phase error versus open loop gain error.

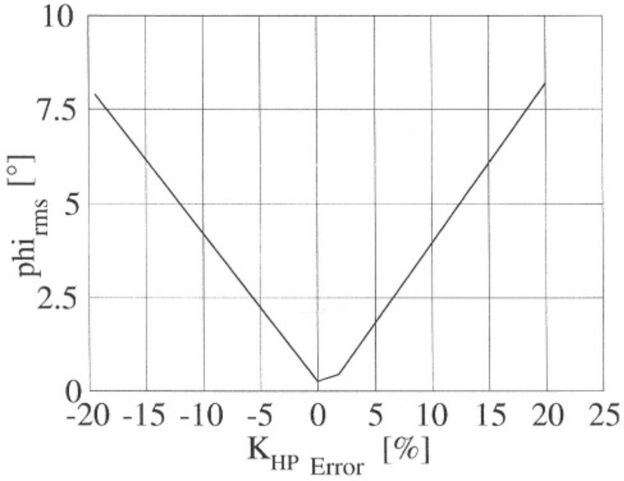


Fig. 14. RMS phase error versus high pass gain error.

8. Conclusion

Fractional-N frequency synthesis has experienced a continuous improvement during the last two decades. The early implementations, which could be classified as fractional-N of the first kind, found no widespread application. The improvement of the first idea, using analogue cancellation schemes, already found more acceptance, especially in measurement equipment. But due to the problems with analogue impairments these fractional-N synthesizers of the second kind still were not the choice for highly integrated mobile phone transceiver ICs. In recent years several research groups were investigating an all digital solution which can be classified as fractional-N of the third kind. This technique has the potential of revolutionizing the synthesizer design. But special care has to be taken, especially in the design of some of the analogue building blocks of the PLL. This frequency synthesizer is of great interest for GSM transceivers, because it offers the possibility to design highly integrated modulators with a minimum number of building blocks.

References

[1] F. de Jager, “Delta Modulation – a Method of PCM Transmission using the one unit Code”, Philips research report, vol. 7, 1952, pp. 442-466.

[2] H. Inose, Y. Yasuda, I. Murakami, “ A Telemetering System by Code Modulation – ΔΣ modulation”, IRE Transactions on Space Electron. Telemetry, vol. SET-8, Sept 1962, pp. 204-209.

[3] W. R. Bennett, “Spectra of Quantized Signals”, Bell Systems Technical Journal, vol. 27, July 1948, pp. 446-472.

[4] G. C. Temes, R. Schreier, S. R. Northworthy, “ Delta Sigma Data Converters; Theory, Design and Simulation”, IEEE Press, 1997.

[5] P. M. Aziz, H. V. Sorensen, J. Van der Spiegel, “An Overview of Sigma-Delta Converters; How a 1-Bit ADC Achieves More Than 16-Bit Resolution”, IEEE Signal Processing Magazine, pp. 61-84.

[6] U. L. Rohde, “Microwave and Wireless Synthesizers; Theory and Design”, John Wiley & Sons, 1997.

[7] P. V. Brennan, “Phase-Locked Loops; Principle and Practice”, MAC-MILLAN Press, 1996.

[8] T. F. Hock, “Synthesizer Design with Detailed Noise Analysis”, RF Design, July 1993, pp. 37-48.

[9] V. F. Kroupa, “Noise Properties of PLL-Systems”, IEEE Transactions on Communications, Vol. com-30, no. 10:2244-2252, 1982.

[10] B. G. Goldberg, “The Evolution and Maturity of Fractional-N PLL Synthesis”, Microwave Journal, September 1996, pp. 124-134.

[11] Sciteq model SEI-1618.

[12] B. Memmler, E. Goetz, G. Schoenleber, “New Fast-Lock PLL for mobile GSM GPRS applications”; ESSCIRC2000, 26th European Solid-State Circuits Conference, Stockholm, Sweden, 19-21 September 2000, pp. 468-471.

[13] R. Best, "Phase Locked Loops – Theory, Design and Applications", McGraw-Hill, 1996.

[14] C. Concling, "Fractional-N Synthesizers Trim Current, Phase Noise", Microwave & RF, February 1998, pp 126-134.

[15] National Semiconductor LMX2350/LMX2352

[16] D. P. Owen, J. N. Wells, "Fractional-N Synthesis System", GEC Review, vol. 11, no. 1, 1996, pp. 3-10.

[17] B. Miller, B. Conley, "A Multiple Modulator Fractional Divider", Transactions of the Forty-Fourth Annual Symposium on Frequency control, 1990, pp. 559-598.

[18] Y. Matsuya, K. Uchimura, A. Iwata, T. Kaneko, "A 17-Bit Oversampling D-to-A Conversion Technology Using Multi Stage Noise Shaping", IEEE Journal of Solid-State Circuits, vol. 29, no. 4, August 1989, pp. 969-975.

[19] G. Märzinger, "Neuartige digitale Modulationsarchitektur mit hoher spektraler Effizienz für Mobilfunkanwendungen", Dissertation, Uni Linz, November 2000.

[20] W. Rhee, B. Bisanti, A. Ali, "An 18-mW 2.5-GHz/900-MHz BiCMOS Dual Frequency Synthesizer With <10-Hz RF Carrier Resolution", IEE Journal of Solid-State Circuits, vol. 37, no.4, April 2002, pp. 515-520.

[21] M. H. Perrott, "Techniques for High Data Rate Modulation and Low Power Operation of Fractional-N Frequency Synthesizers", PhD thesis, Massachusetts Institute of Technology, 1997.

[22] T. A. D. Riley, M. A. Copeland, "A Simplified Continuous Phase Modulator Technique", IEEE Transactions on Circuits and Systems – II: Analog and Digital Signal Processing, vol. 41, no.5, pp. 321-328.

[23] Technical Specifications GSM 05.04. and GSM 05.05.

[24] S. Heinen, S. Herzinger, "Transmitter Concepts, Integration and design trade-offs", Circuits and Systems for Wireless Communications, Kluwer, 2000, pp.141-155.

[25] B. Neurauter, G. Märzinger, T. Lüftner, R. Weigel, M. Scholz, V. Mutlu, J. Fenk, "Low Phase Noise Frequency Synthesizer with Very High Frequency-Resolution for GSM900/DCS1800", Proc. 2000 European Conference on Wireless Technology, Oct 2000, Paris, France, pp. 29-33.

[26] B. Neurauter , G. Märzinger , A. Schwarz , R. Vuketich, M. Scholz, R. Weigel, J. Fenk, „GSM 900/DCS 1800 Fractional-N Modulator with Two-Point-Modulation ", IEEE RFIC Symposium 2002, Seattle, pp. 425-428.

This page intentionally left blank

PART II: DESIGN FOR ROBUSTNESS

The trends towards deep submicron technologies, higher operation frequencies and harder application environments, results more and more in integrated design robustness issues. ESD (Electrostatic Discharge), EMC (Electromagnetic Compatibility) EMS (Electromagnetic Susceptibility), EME (Electromagnetic Emission) are all unfamiliar terms for the classical IC designer. However those issues are becoming major concerns for integrated circuit designers the last few years. As a result, the understanding and the in-calculation in the design of those effects (the so called co-design) are becoming indispensable.

The first three papers are addressing the ESD issue. The first paper describes the problems and provides the necessary information for analog designers to understand the problem. Some solutions depending on the application and specification are addressed.

The second paper focuses the ESD issue more on ‘smart power’ devices. Since power output drivers are very often used in very aggressive environments, they require very high ESD performances.

The third paper describes the co-design in low noise amplifiers (LNA) for radio frequencies. Classical ESD topologies immediately destroy the noise performances, and as a result careful co-design with ESD topologies and LNA structures is required.

The next two papers (papers four and five) deal with EMC issues. The fourth paper discusses the analysis and test verifications for EMC specifications. It is shown that as a result of it, ‘hot spots’ can be located, which is of course valuable information for the designers.

The fifth paper addresses the design issues related to EMC. Design flows of the different topics (EME, EMS, ESD) are addressed and it is clear that from the beginning of the design cycle those important issues have to be in calculated to achieve high quality robustness industrial designs.

The last paper deals with the electrical signal coupling mechanisms on the chip and connecting the circuits. Extreme care in incorporating the knowledge of bonding schemes, route and placing is as a result utmost important to end up in robust mixed-mode circuit designs and avoiding EMC problems. Different tools, strategies and design techniques are addressed and analyzed.

Michiel Steyaert

This page intentionally left blank

ESD FOR ANALOGUE CIRCUIT DESIGN

Dave Clarke (1) and Alan Righter (2).

(1) Analog Devices, Raheen Business Park, Limerick, Ireland.
(2) Analog Devices, 804 Woburn Str. Wilmington, Mass, USA.

Abstract

Designing for Electrostatic Discharge (ESD) robustness is a critical requirement for Integrated Circuits (ICs) since ESD voltages as low as the Gate Oxide breakdown can damage an IC. This is especially true for the newer lower Geometry Processes where the Gate Oxide breakdowns are reducing all the time. ESD Protection for Analogue or mixed signal parts is an even greater challenge due to the various requirements of analogue pins e.g. low capacitance, low leakage, high voltage etc. Typically this requires the development of specific ESD Cells for the various pin applications. This paper covers overall Protection schemes for both Digital & Analogue parts, with particular focus on the ESD requirements for Analogue pins and some case studies to hi-light some of the issues typically encountered with ESD Design.

1. Introduction

Electrostatic Discharge (ESD) is the transient discharge of static charge, which can occur due to Human handling, Machine handling or charge build-up due to friction.

This ESD action occurs in 3 steps;

1) Friction between 2 dissimilar materials may cause static charge to build up, with different polarities on each.
2) Separation of materials causes 2 charged objects of opposing polarities.
3) Discharge may occur due to direct contact or by arcing.

The polarity of the ESD pulse may be positive or negative depending on the polarity of the charged object.

ESD pulses as low as the Gate Oxide breakdown voltage (BV) rating of the Fabrication Process of the IC may cause Latent or Catastrophic damage to the silicon for MOS based technologies.

A. van Roermund et al. (eds.), Analog Circuit Design, 131-168.
© 2003 *Kluwer Academic Publishers. Printed in the Netherlands.*

Everyday actions can generate significant ESD voltages. If you feel a zap you are talking about voltages in excess of 3 kV. If you see a spark due to arcing, you are talking about voltages in excess of 6 kV.

Therefore the 2 techniques to protect ICs from ESD damage, is to (1) minimise exposure to ESD by creating a static aware environment i.e. using dissipative materials, ESD wrist-straps, Grounded surfaces etc. (2) improve the ESD robustness of the IC itself.

2. ESD Models

There are 3 primary models to simulate real world ESD mechanisms;

1) Human Body Model (HBM) to simulate a charged person discharging to an IC [1], [4]. A high voltage supply charges up the Charging Capacitor (C1) through the Power Supply series resistor (R1). The equivalent circuit for HBM ESD is a 100 pF capacitor discharging through a 1.5 Kohm (R2) series resistor as per the circuit in Fig 1.

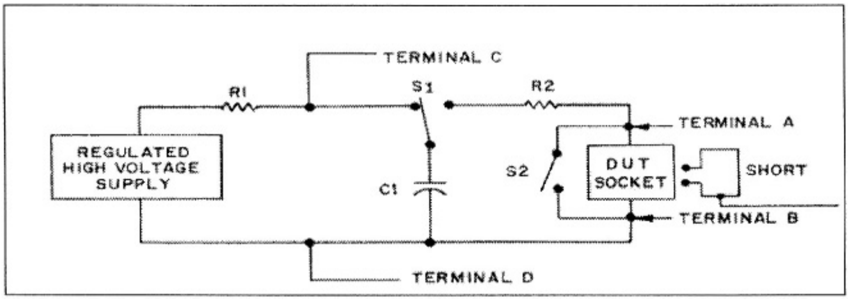


Fig. 1. ***HBM & MM ESD Tester equivalent circuits***

The waveform through a short circuit is a double exponential with <10 nS rise time and <150 ns time constant for the Fall time as per Fig 2.

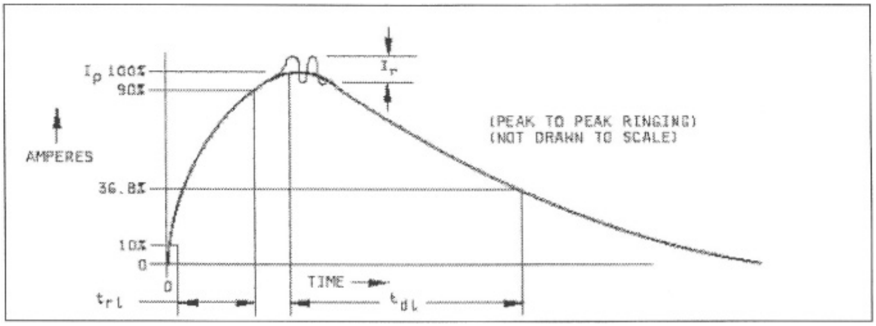


Fig. 2. HBM Waveform

2) Machine Model (MM) to simulate a discharge to an IC through a conductor [2]. The equivalent circuit is the same as HBM except that the charging capacitor is 200 pF through zero ohms series resistance. This model correlates to HBM and is sometimes called the ‘Zero-Ohm Model’ or ‘Worstcase HBM’ since it correlates to HBM with typically a 10-20x correlation i.e. 100 V MM = 1-2 kV HBM. Obviously low impedance pins are more sensitive to MM ESD. The ESD damage seen for MM is typically similar to HBM except that the extent of the damage is far more severe e.g. cratering, metal tracks blown open circuit.

3) Charge Device Model (CDM) to simulate discharge to Gnd of a charged IC, which was charged due to friction e.g. IC travelling through a handler chute, or charged due to an induced field [3]. Robotic CDM is where a robotic arm charges all pins to the same voltage before each pin is discharged, while Field Induced CDM is where the DUT is charged capacitively through a charging plate as per Fig 3.

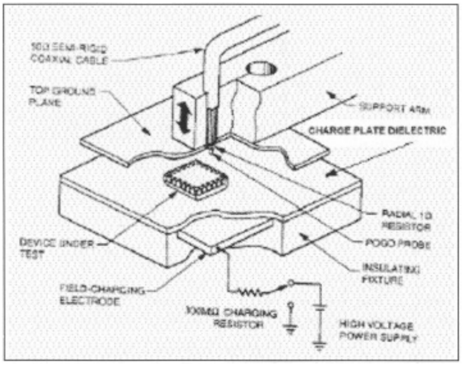

50Ω SEMI-RIGID COAXIAL CABLE
TOP GROUND PLANE
SUPPORT ARM
CHARGE PLATE DIELECTRIC
RADIAL 1Ω RESISTOR
POGO PROBE
INSULATING FIXTURE
DEVICE UNDER TEST
FIELD-CHARGING ELECTRODE
100MΩ CHARGING RESISTOR
HIGH VOLTAGE POWER SUPPLY

Fig.3. Field Induced CDM ESD Tester

The charging capacitor for CDM is therefore the package itself i.e. variable capacitance, therefore CDM, unlike HBM or MM, is package dependent. Discharge occurs through a 1 ohm series resistor to Ground.

The CDM waveform is a damped sinusoid with <400 ps rise-time for the initial peak as per Fig 4.

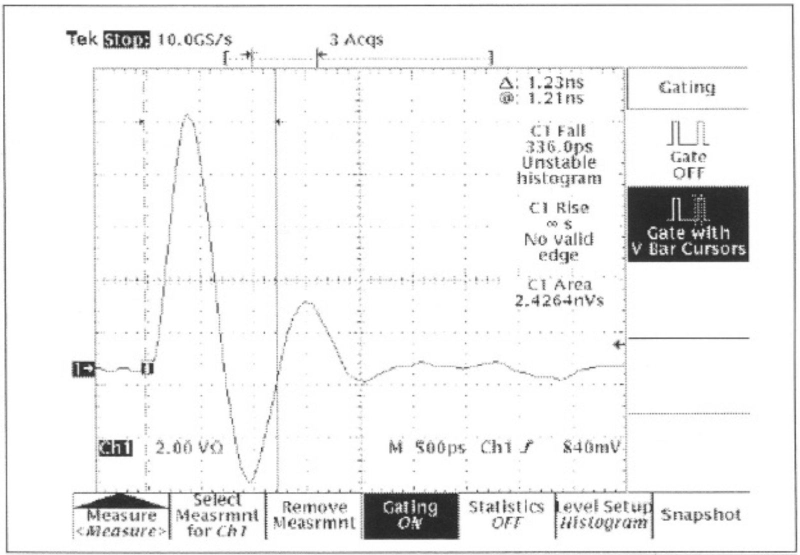
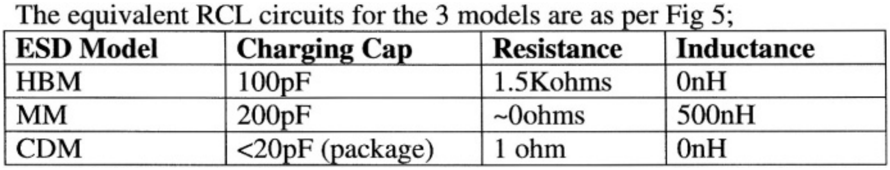

Tek Stop: 10.0GS/s
3 Acqs
Δ: 1.23ns
@: 1.21ns
C1 Fall
336.0ps
Unstable
histogram
C1 Rise
∞ s
No valid
edge
C1 Area
2.4264nVs
Gating
Gate
OFF
Gate with
V Bar Cursors
Ch1 2.00 VΩ
M 500ps Ch1 ∫ 840mV
Measure
<Measure>
Select
Measrmnt
for Ch1
Remove
Measrmnt
Gating
ON
Statistics
OFF
Level Setup
Histogram
Snapshot

Fig. 4. CDM Waveform

Even though the current levels are high, the short duration of the pulse means that the energy dissipated is low, therefore the typical damage location is Gate Oxide, either catastrophic or latent i.e. failure mechanism is recoverable.



The equivalent RCL circuits for the 3 models are as per Fig 5;

ESD Model	Charging Cap	Resistance	Inductance
HBM	100pF	1.5Kohms	0nH
MM	200pF	~0ohms	500nH
CDM	<20pF (package)	1 ohm	0nH

Fig. 5. ESD Model Equivalent RLC circuits

3. Analogue ESD Protection vs Digital ESD Protection

The major difference between ESD Protection Design for Analogue vs Digital parts is in the Input/Output (I/O) Pad Protection. The same techniques can be used for Power Supply Protection and overall Protection schemes but the functionality, operational specs and/or attached circuitry can make Analogue parts far more difficult to protect.

Some typical examples of analogue pin requirements are;

- High Voltage operation (outside the Power Supply Rails)
- Low Leakage Performance
- High ESD Ratings e.g. 15 kV IEC for Interface or Touch-screen pins
- Low Capacitance e.g. RF or high frequency applications
- Low input resistance

Being limited to low Input resistance can make ESD protection of Analogue pins very difficult, as it makes it far more difficult to isolate the internal circuitry from the ESD pulse.

Analogue parts may also use different technologies to the standard CMOS for Digital, for example;

- BiCMOS
- BiMOS
- Bipolar
- DMOS (Smart Power)
- SOI

4. Design for ESD Protection techniques

The standard ESD Protection scheme for I/O pins is using the “H Network” protection scheme, which uses 4 ESD Cells as per Fig 6.

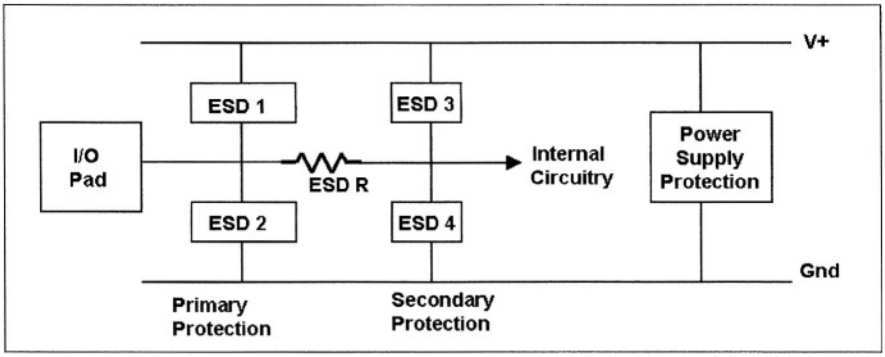


Fig.6. ESD “H Network” for I/O Pad Protection

Protection Devices 1 & 2 provide Primary Protection i.e. they shunt most of the ESD current to the Power Rails. The ESD Resistor R isolates the internal circuitry from the pad. Protection Devices 3 & 4 provide secondary protection, they provide voltage clamps inside the ESD resistor.

Even if the Primary Protection clamps the voltage at the pad, the pad voltage may rise to dangerous levels for the sensitive internal circuitry due to the on resistance (Ron) of the Primary Protection. The secondary protection allows local clamping near the internal circuitry at a lower voltage than the pad. This voltage difference is dropped across the ESD resistor R. The secondary protection may be a scaled down version of the primary protection or may be a different ESD device with a lower turn-on voltage. If the secondary protection turns on 1st, then the voltage drop across the ESD R will eventually force the primary protection to turn on.

There are 4 possible ESD pulse combinations at an I/O;

1) Positive ESD pulse from pad to V+. Primary Protection 1 will usually forward bias with secondary clamping from Protection Device 3.
2) Negative ESD pulse from pad to V-. Primary Protection 2 will usually forward bias with secondary clamping from Protection Device 4.

3) Positive ESD pulse from pad to Ground. Primary Protection 1 will usually forward bias with secondary clamping from Protection Device 3, bringing up the V+ potential to turn-on the Power Supply Protection to Ground. If Protection Device 2 is a snapback structure then it may snapback to Ground.
4) Negative ESD pulse from pad to V+. Primary Protection 2 will usually forward bias with secondary clamping from Protection Device 4, bringing down the Ground potential to turn-on the Power Supply Protection to V+. If Protection Device 1 is a snapback structure then it may snapback to V+.

Therefore the importance of the Power Supply Protection is not just for ESD pulses across the rails but may also be critical as the ESD path for I/O pulses.

5. Ideal ESD Structure

The ideal ESD Structure should have 2 properties;
1) Always Electrically open during normal operation (i.e. infinite resistance, zero capacitance & zero inductance)
2) Instantaneously electrically closed during an ESD event (i.e. zero resistance, zero capacitance & zero inductance)

Due to the laws of physics we cannot get to this ideal as all ESD structures will have parasitic leakage, capacitance & inductance that will affect the normal operation of the pin. During an ESD event, the ESD structure will have a finite turn-on response time and will have finite resistance.

This gives the 3 golden rules for ESD Protection Design;
ESD Structure should
1) Turn-on before any sensitive internal circuitry
2) Be able to shunt ESD current away from sensitive internal circuitry i.e. be able to handle Amps of ESD current
3) Have minimum effect on normal circuit operation

Typically the ESD structure is required to be as large as possible to be able to handle the ESD current & reduce the series Ron. Increasing the size of the ESD structure obviously increases the capacitance of the structure also, which helps from an ESD point of view i.e. increased capacitance at pad to dampen ESD pulse rise-time, but this increased capacitance may also interfere with the performance of part/pin in normal operation. Parasitic leakage is also a concern, especially at temperature.

There is also the issue of silicon area, having a large ESD Cell with fantastic ESD performance is a luxury that may not be available in an environment of cost efficiency, therefore there is always a trade-off between acceptable ESD performance vs essential part performance and with a constant focus on Die area reduction.

6. CMOS ESD Protection

Typical ESD Structures used for CMOS ESD Protection are Off-NMOS, Off-PMOS and triggered NMOS. The basic ESD NMOS device is usually with the Drain tied to the pad and the Source & Gate tied to Ground i.e. Grounded Gate NMOS (ggNMOS). This device has a parasitic NPN device as can be seen from Fig 7. The Drain forms the Collector, the Backgate forms the Base & the Source forms the Emitter.

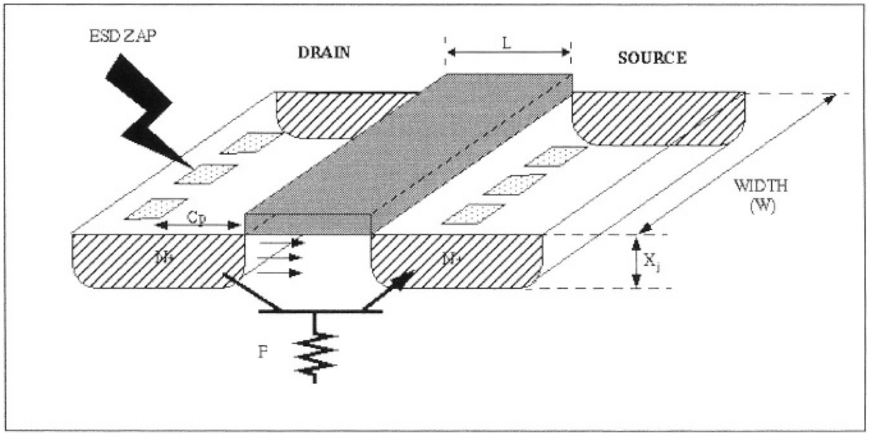


Fig.7. ESD NMOS Cross-section

When the Stress voltage at the Drain is increased to the avalanche BV of the N+ to P-substrate junction, hole electron pairs arc formed. This hole current is collected by the substrate pickup generating a voltage gradient across the substrate resistance. If this voltage gradient reaches the Vbe voltage anywhere along the P-/N+ Source junction, then the junction is forward bias and injects electron current into the substrate which adds to the Drain current. At this stage the parasitic is turned on in self-biased mode i.e. Drain N+/P- avalanche BV generates the Base current.

The turn-on of this parasitic NPN causes a negative resistance region as more carriers are available for multiplication until a minimum voltage is reached called the holding voltage.

This mechanism is called snapback and is a key requirement for snapback based ESD Protection, by allowing a stress voltage to cause a snapback to a lower voltage (i.e. voltage clamping).

This snapback characteristic as shown in Fig 8, with the trigger voltage at It1 is non-destructive. But the 2nd snapback at It2 is a thermal mechanism and therefore is destructive.

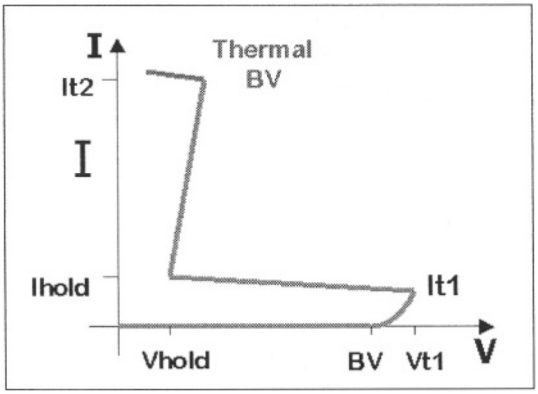


Fig.8. Snapback Characteristic

A key objective in ESD Cell design is to increase this It2 point and therefore increase the current density capability of the structure. This can be done with careful layout optimised for ESD.

There are 3 critical Design Rules for the layout;

1) Gate Length L, this determines the Base width of the parasitic bipolar & it's gain. Typically the Gate L will be at or close to the geometry minimum.

2) Gate Width W. The junction depth (Xj in Fig 7) determines the sidewall area on the Drain junction for a particular Gate W, which also determines the ESD Current density. The source/drain junction depth Xj is a process parameter and therefore cannot be changed easily but the Gate W can be increased in layout. The ESD performance correlates directly to the Gate W provided the full Gate W is effective.

3) The Drain Contact to Poly Gate spacing (Cp in Fig 7) is an ESD spacing to keep the Drain contacts away from the hot drain sidewall during the ESD

event. It also adds ballast resistance to the Drain giving more uniform turn-on across the full channel width, both within fingers and across fingers [11].

If the Source is shorted to the backgate than a negative pulse causes from the Drain to the Source causes the parasitic diode from the P backgate to the N Drain to forward bias, therefore no ESD spacing is required on the Source side. But if the Source is isolated from the backgate, (e.g. for noise isolation purposes) then ESD spacings are required on the both the Drain and Source sides.

Using a single fingered ESD device will give non-uniform turn-on and therefore inefficient use of the channel width. But using a multi-finger with a maximum finger length defined for the Process allows more efficient use of the full channel width as per Fig 9. In reality each finger snaps back separately. When the drain voltage reaches Vt1, the 1st finger snaps back to Vhold. As the ESD stress is increased, the drain voltage rises due to the Ron of the 1st finger and the 2nd finger snaps back. This process continues until hopefully all the fingers have turned on, resulting in a sawtooth characteristic in the snapback curve. Therefore it's critical that the trigger voltage is less than the Thermal breakdown voltage [8] i.e. Vt1 < Vt2. It's also critical that the length of the ESD finger is large enough so that a single finger can handle the ESD current before the 2nd finger snaps back. This leads to an ESD design rule for the minimum finger length. Typical range for these min/max finger lengths are 15-100 um, with smaller design rules for the lower geometries [5], [6]. These ESD Design Rules need to be determined on testchip for a particular Process. Getting uniform turn-on of all the ESD fingers is difficult [7], [8].

This multi-finger layer technique should also be used for large Diode or Bipolar ESD devices.

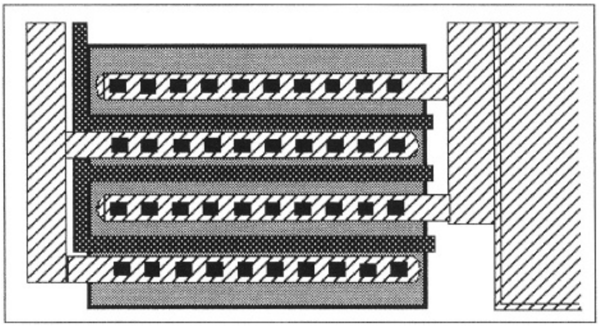


Fig.9. Anti-parallel Finger or Ladder Layout of ESD MOS device

By applying a Gate voltage to the NMOS device the trigger voltage can be reduced below the avalanche BV of the N+/P- junction [9], [7], [8] as per Fig 10. This is an ideal way to reduce Vt1 below Vt2 and therefore improve the multi-finger turn-on performance. The reduced Vt1 occurs due to the Gate voltage going above the Threshold voltage causing an increase in the substrate current (Isub) i.e. MOS conduction and this allows the It1 point to occur at a lower voltage. Ideally the peak of the bell-shaped curve (Isubmax) is the ideal bias to apply to the Gate.

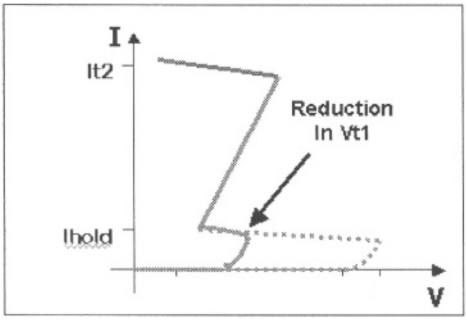


Fig.10. Snapback Characteristic with Gate Bias reducing Vt1

During an ESD event there is typically coupling from the Drain diffusion to the Gate due to the Drain to Gate Capacitance. This can be improved upon by adding extra capacitance between the Gate and Drain terminals [9], [10] as per Fig 11. The resistor is required to pull the Gate back to Ground after the ESD event. This RC effect on the substrate current (Isub) can be simulated in SPICE, to optimise it’s performance.

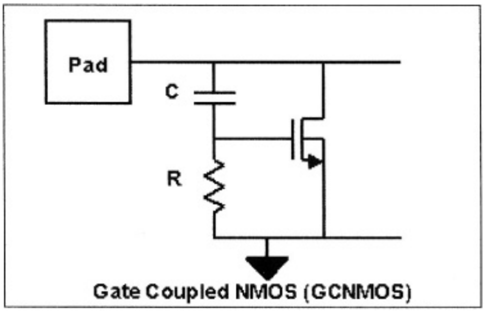


Fig.11. Gate Coupled ESD NMOS (GCNMOS)

An extension of this technique is to use Gate Driven NMOS (GDNMOS) whereby the ESD NMOS gate is driven by an ESD sensing circuit [12], [13], [14].

For CMOS or BiCMOS ESD Protection, the standard H network approach can be used as per Fig 12;

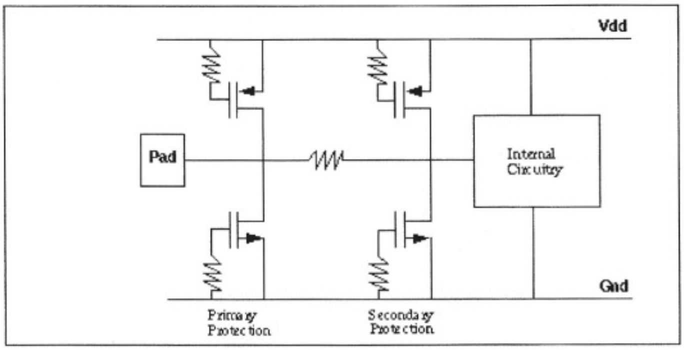

Vdd
Pad
Internal Circuitry
Gnd
Primary Protection
Secondary Protection

Fig.12. CMOS Input Protection

The secondary protection is a scaled down version of the primary Off-NMOS/PMOS protection and layout wise can be incorporated as 1 finger of the primary ESD protection.

Analog Devices (ADI) uses a patented timer circuit that uses the ESD pulse to drive the ESD NMOS gate to 1-2V during the ESD event [15]. This Gate driven ESD NMOS device is used for Power Supply Protection as per Fig 13;

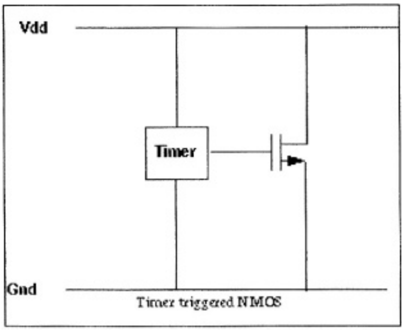

Vdd
Timer
Gnd
Timer triggered NMOS

Fig.13. Power Supply triggered NMOS

Similar Grounds Rails are protected with respect to one another with Back-to-back diodes as per Fig 14. If a part has a large number of Grounds, an ESD Ground may be used as a reference whereby all Grounds are diode clamped to the ESD Ground. This means that there is no more than a 2 diode drop in the ESD path between any pair of Grounds.

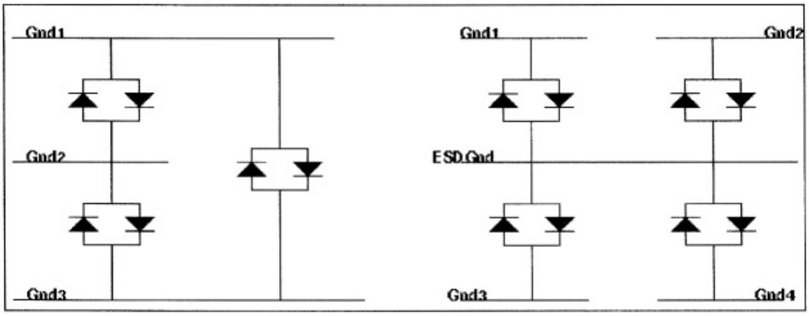


Fig.14. ESD protection for separate Ground signals

Power Supplies of the same potential can also be diode clamped in the same fashion as the Grounds. This isn't essential as the Power Supply Protection Cells provide a path through the Ground Clamps.

MOS Output drivers can be made self protecting by giving them ESD spacings, but if the Gate L's are not adequate, the required ESD Gate L can be made up with extra fingers of an Off-NMOS/PMOS as per Fig 15. A small ESD resistor can be added in series with the Output Drivers, if the preferred ESD path is through the ESD devices only.

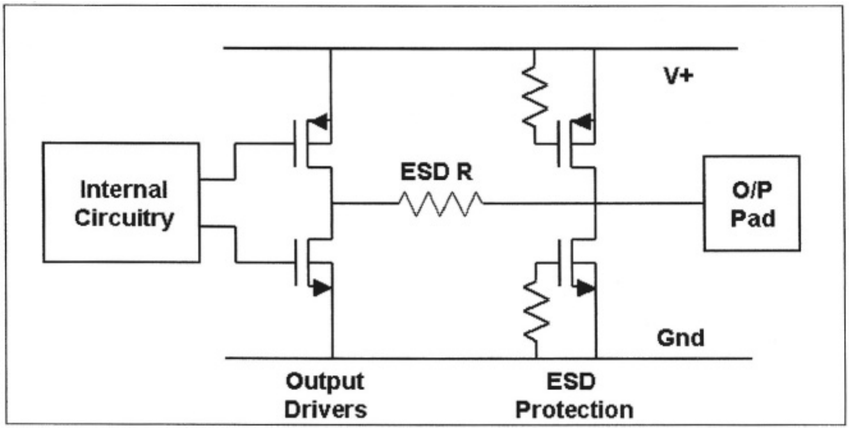

V+
Internal Circuitry
ESD R
O/P Pad
Gnd
Output Drivers
ESD Protection

Fig.15. MOS Output Driver ESD Protection

In deep sub-micron processes the floating gate O/P driver may have the advantage in snapping back 1st i.e. part is un-powered, therefore soft-pull devices can be used to tie the gates of the ESD devices off, rather than tying the gates hard off through a resistor [16] as per Fig 16. This makes the ESD & O/P drivers devices comparable during the ESD event i.e. gates floating.

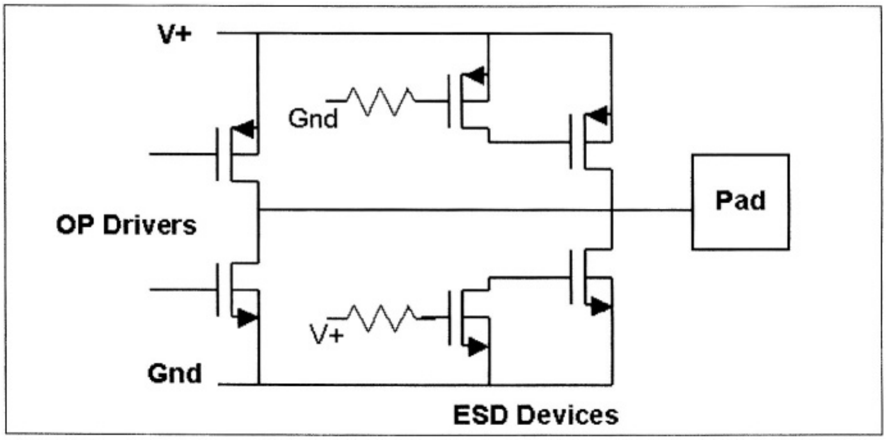

V+
Gnd
Pad
OP Drivers
V+
Gnd
ESD Devices

Fig.16. Output Driver Protection with Softpull ESD devices

For Top Level ESD Protection on a part with multiple Power Supply blocks, the scheme in Big 17 can be used. It's critical that each Power Supply block has it's own Power Supply Protection Clamp. Protection between blocks typically occurs through the Gnd pins, which is typically substrate for sub-micron processes. The clamping between Grounds is back-to-back or anti-parallel diodes, either directly to all other Grounds or indirectly through an ESDGnd rail as per Fig 14. Diode or MOS clamping between the Power Supplies can also be included but not be allowed due to the voltage ranges or Power-up sequence concerns.

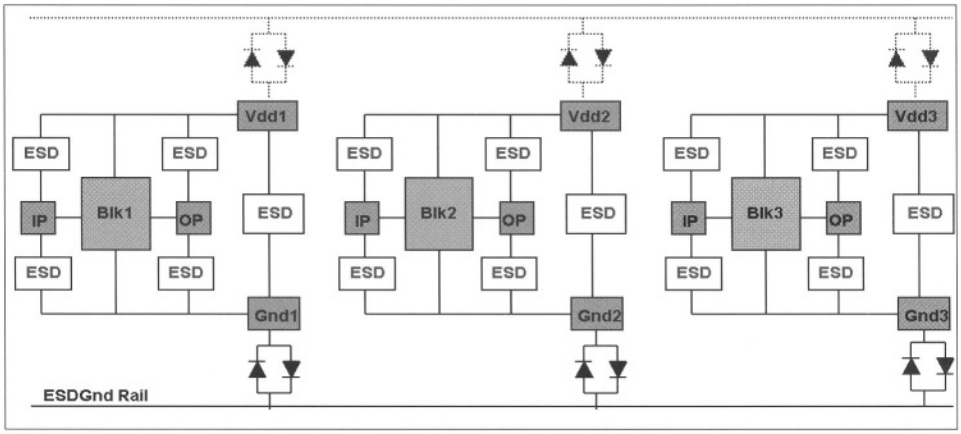


Fig.17. Top Level ESD Protection Scheme

7. Analogue Pin Protection

Some examples of Analogue pins requiring unusual ESD Protection are High Voltage pins operating outside the rails, low Capacitance pins for High Frequency or RF applications and Bipolar inputs.

7.1. HV pin ESD Protection

Analogue parts can have pins that operate outside the rails, for example a 5V ADC with +/-10V Ain.

These pins typically go into internal attenuator circuits where the high external voltages are attenuated to the voltage levels of the internal circuitry. These attenuators typically consist of high value resistor strings e.g. Thin Film (TF) or Polysilicon Resistors.

Since these pins can operate at voltages outside the rails, we cannot use diode or parasitic diode clamping to the rails, therefore HV ESD structures are required which voltage clamp with respect to one of the rails.

Typical structures that can be used are based on;

1) NMOS
2) SCR
3) Bipolar, either BE shorted or triggered with an ESD sense circuit
4) Floating Base Bipolars, if bi-directional breakdown voltages are required

If the voltage range of the Ain pin is outside the highest junction BV's of the Process then the only option is to make the Attenuator resistor string itself ESD robust i.e. self-protecting.

For TF (or Poly) resistors, the heat dissipation properties are very poor because they are isolated between oxide layers, therefore the Width (W) of the resistor is critical to ensure that the resistor can handle the ESD current density. The maximum ESD current density of the TF needs to known, so that the resistor width W can be determined, taking into account the 1.5 Kohm HBM series R as per Fig 18.

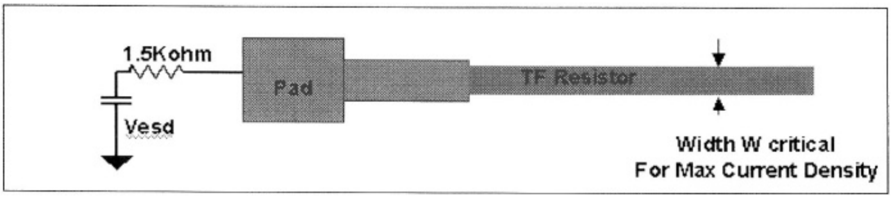


Fig.18. Unprotected TF ESD

If there is no voltage clamp at the pad, the pad voltage can reach excessive levels, therefore the Bondpad structure itself may breakdown first i.e. underlying oxide to substrate, especially if the pad structure contains a Polysilicon layer. Therefore a top Metal layer/layers pad should be used to increase this pad breakdown voltage as per Fig 19. Obviously these Metal only pads need to be qualified to verify that there are no Assembly Reliability issues. It is also important that these unprotected tracks don't cross any other tracks, as you may zap across the oxides between tracks.

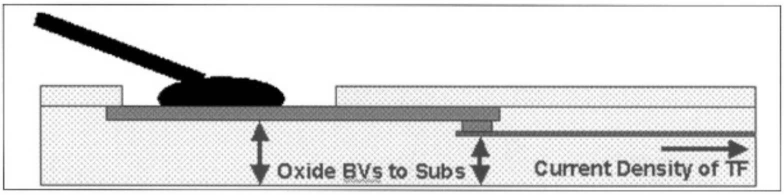


Fig.19. Cross section of unprotected TF pad

For a multifinger TF resistor, the layout is critical. There should be no corners in the resistor, therefore the resistor should be in multiple stripes with connections between stripes on Metal (with rounded corners). The exit point from the resistor should be away from the pad entry point as per Fig 20.

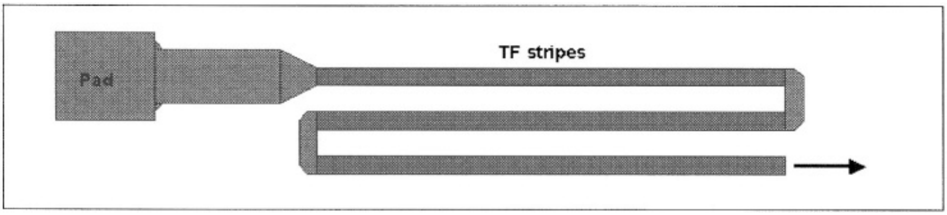


Fig.20. Layout of multi-finger TF Resistor

With these layout techniques it is possible to get >4 kV HBM ESD performance from self-protecting TF resistors. This same protection technique also applies to Poly resistors.

7.2. Low Capacitance ESD

Low Capacitance ESD Structures are essential for High Frequency applications e.g. RF. The ideal solution is to have very small low capacitance ESD structures that will give the required ESD performance and rely on excellent Power Supply Protection circuitry to minimise the ESD stress on the RF ESD Protection. The simplest RF ESD protection scheme is small RF diodes to the rails as per fig 21.

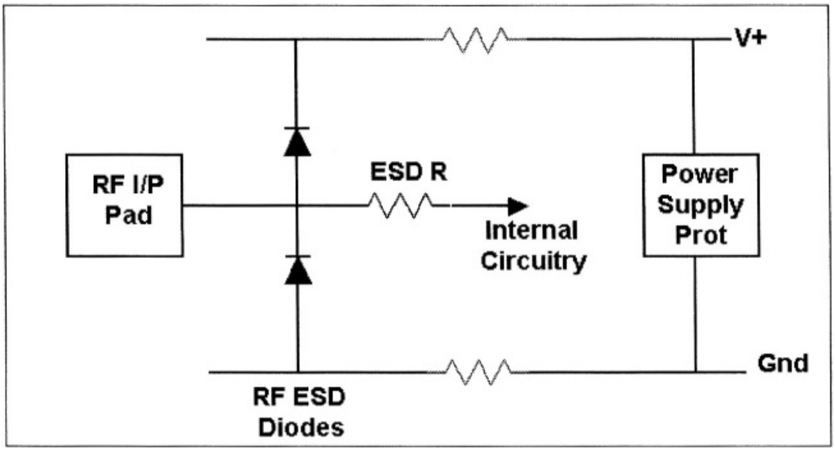


Fig.21. RF ESD Protection schematic

It is critical that the Power Supply Protection is as close as possible to the RF Pad to minimise the chance of the small RF diodes going into reverse breakdown due to the parasitic voltage drops across the Power Supply rails e.g. positive ESD pulse at pad to Ground or negative ESD pulse at pad to V+. Since the low capacitance RF diodes are so small, they are very sensitive to ESD damage in reverse breakdown. The ESD resistor needs to be as large as can be tolerated by the application.

Using this RF ESD Protection scheme, with RF diodes in the 60-120 fF range for >7 GHz operation, it's possible to get >2 kV HBM ESD.

Another option is to only have protection with respect to one of the rails, usually Ground/Substrate, as this reduces the Capacitance further.

If very low capacitance isn't critical then scaled down Off-NMOS devices can be used. Off-NMOS devices with scaled down Gate W (<120um) with reduced ESD spacings (Drain contact to Source) are capable of achieving 2 kV performance. Obviously these reduced ESD spacings need to be determined on testchip for their ESD performance.

These RF diodes could also be replaced with SCRs (Silicon Controlled Rectifiers), as the SCR is the most space efficient ESD device available i.e. it has a very strong snapback mechanism with low Ron (1-2ohms typically). The SCR is described in a later chapter.

7.3. Bipolar Protection

XFCB are Xtra Fast Complimentary Bipolar Processes, therefore similar to RF applications, low capacitance is critical. ADI use patented circular diodes [17] for low voltage XFCB Processes. Circular diodes give uniform field intensity across the full channel width and eliminate the issues of increased field intensity at the ends of fingers. ESD Cell capacitances are as low as 30fF, with ESD performances of up to 4 kV.

Fig 22 shows a picture of a pair of the XFCB diodes used on an I/O pad to provide diode clamping to the rails.

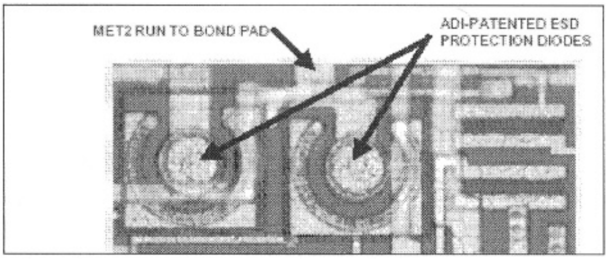


Fig.22. Photo of XFCB circular ESD diodes

For Bipolar Inputs 100ohms series resistor is recommended to protect Bipolar Bases, with back-to-back diode clamps between differential inputs, stacked if required by input voltage swing.

Power Supply Protection can be reversed bias ESD diodes in parallel (to reduce Ron) or Bipolar based snapback structures. Power Supplies and Grounds are diode clamped with back-to-back diodes where possible.

8. Transmission Line Pulse Tester

The Transmission Line Pulse (TLP) Tester is a essential tool for ESD Design [19], [20]. It uses a transmission line to generate a square wave pulse, whose energy level correlates to that of the HBM waveform. The TLP Tester block diagram is shown in Fig 23, where a High Voltage supply charges up the transmission line. The charged transmission line is then switched through an attenuator to a matched impedance scope cable with the current & voltage monitored. The transmission line reflections cause squares waves to be generated whose pulse widths are determined by the transmission line length.

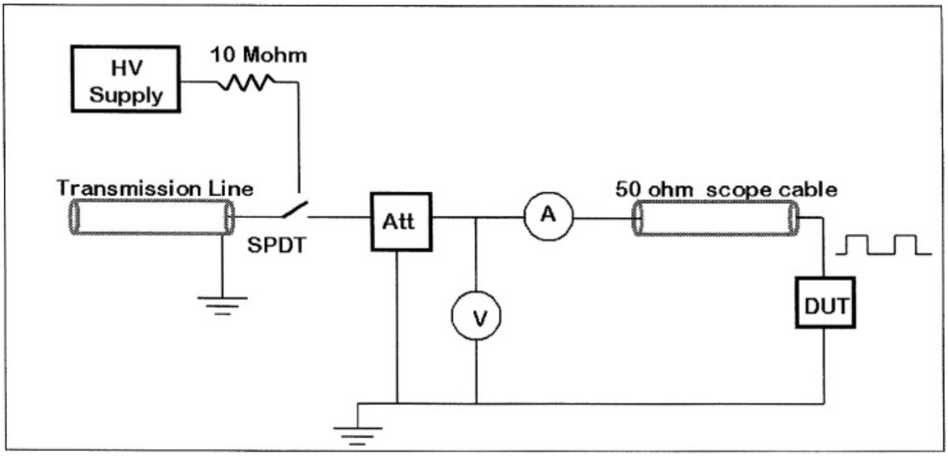

HV Supply
10 Mohm
Transmission Line
SPDT
Att
A
V
50 ohm scope cable
DUT

Fig.23. Constant Impedance TLP Tester Block Diagram

The key advantage of the TLP tester is that it is far easier to generate a repeatable squarewave waveform rather than the double exponential HBM waveform, especially at low voltages (sub 50V level).

The TLP tester measures the current & voltages pulses applied to the DUT [18] as per Figs 24 & 25.

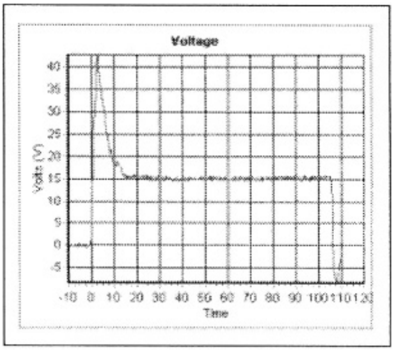


Fig.24. TLP Voltage Pulse

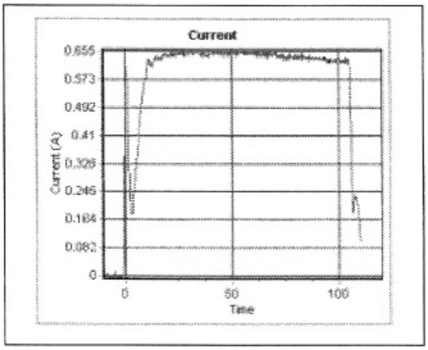


Fig.25. TLP Current Pulse

A series of increasing pulses are produced so that an IV plot can be generated showing the ESD IV characteristic of the Device under Test (DUT) as per Fig 26. After each TLP pulse is applied a leakage measurement is made to detect damage to the DUT, as this fail point may not always be visible from the TLP Plot. This is shown as the vertical series of data points on the left hand side of the TLP plot in Fig 26.

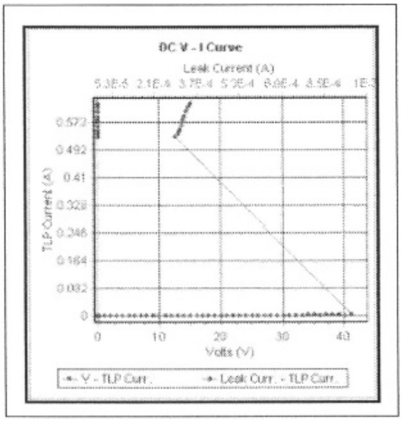


Fig.26. TLP IV characteristic showing snapback

9. Silicon Controlled Rectifier (SCR)

The SCR consists of a PNPN structure whereby you have a PNP to V+ and an NPN to Ground as shown in Fig 27. Once one of the bipolars is triggered then the other bipolar is automatically turned once a Vbe voltage drop occurs across the BE resistor.

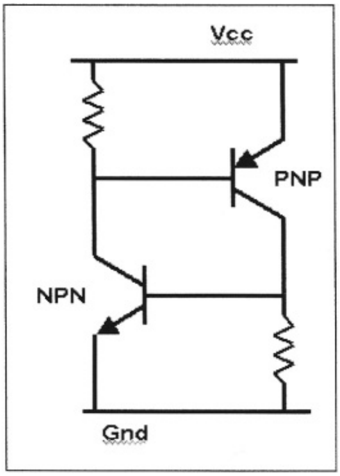


Fig.27. SCR schematic

A Cross-section of the standard SCR [21], [22] is shown in Fig 28. The Anode connection tied to the pad connects to both P+ & N+ diffusions in the N-Well. The Cathode connection tied to Ground connects to both P+ & N+ in substrate. This generates the parasitic vertical PNP and lateral NPN devices with the Well and substrate providing the required BE resistors. This Lateral SCR (LSCR) is triggered by the N-Well to P-Substrate avalanche breakdown and is therefore only suitable for high voltage pins.

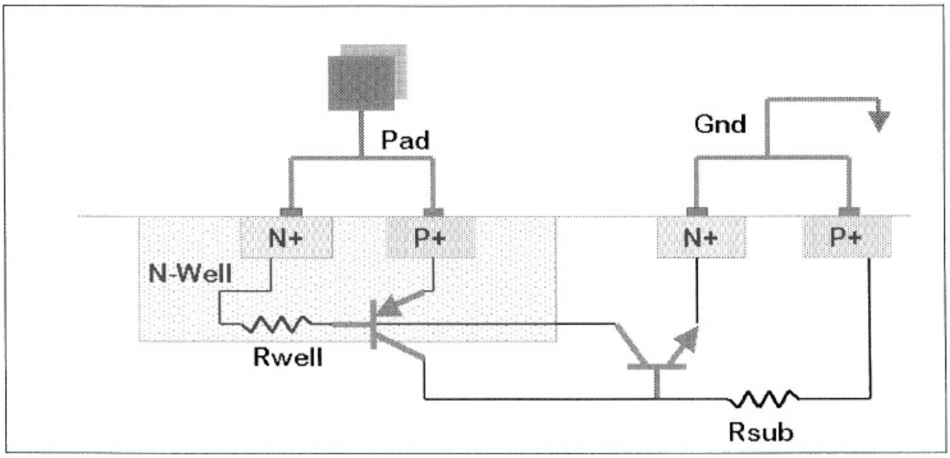


Fig.28. Lateral SCR (LSCR)

In order to get a lower trigger voltage, an N+ trigger diffusion can be added to the N-Well to P-Substrate junction to reduce the trigger voltage to the avalanche BV of N+ to substrate [23] as per Fig 29. This version is called a Modified Lateral SCR (MLSCR).

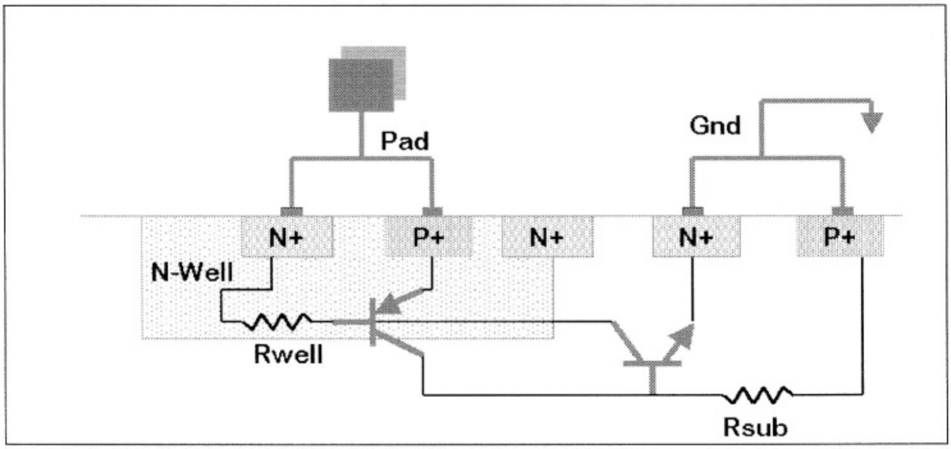


Fig. 29. Modified Lateral SCR (MLSCR)

By adding a Polysilicon gate to the region between the N+ trigger diffusion and the N+ Cathode connection, the SCR changes to a Low Voltage Triggered SCR (LVTSCR), which is now triggered by the Off-NMOS device [23],[24] as per Fig 30.

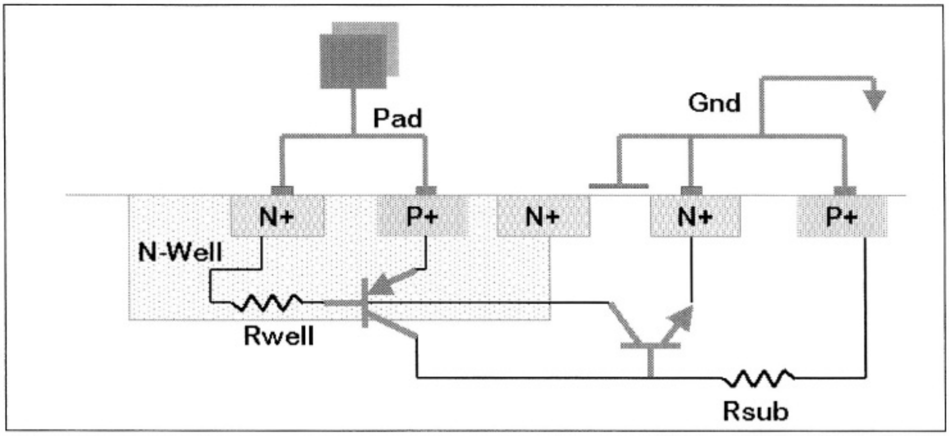

Pad
Gnd
N+
P+
N+
N+
P+
N-Well
Rwell
Rsub

Fig.30. Low Voltage Triggered SCR (LVTSCR)

The SCR has a very strong snapback mechanism with low Ron after snapback as per TLP plot in Fig 31.

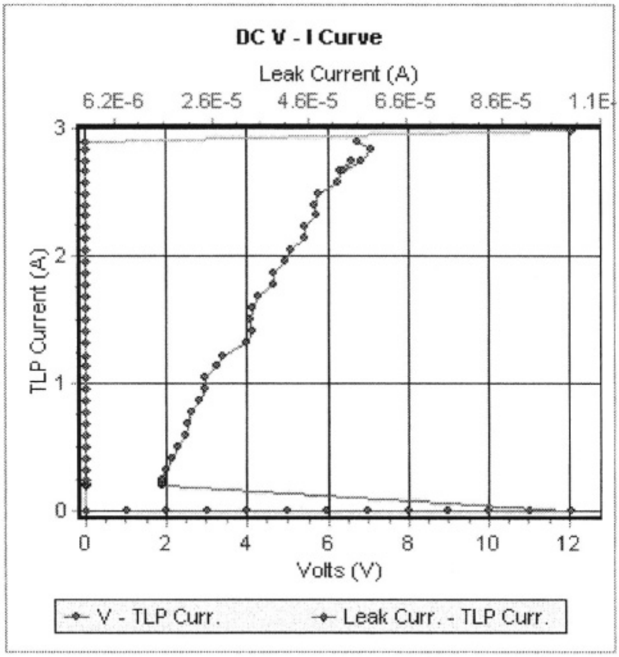


Fig.31. TLP Curve of ESD SCR

Most designers, especially those who have experienced Latch-up issues in CMOS technology, would be very wary of deliberately using SCRs for ESD. But if properly designed an ESD SCR can be controlled. The key parameters that describe the SCR characteristic are, as per Fig 32;

1) Breakdown Voltage (BV), this is the voltage at which current starts to flow, usually due to avalanche.
2) Trigger Voltage (Vt1, Vtrig), this is the voltage at which snapback occurs i.e. BE of Bipolars reached.
3) Trigger Current (It1), this is the current required to trigger snapback.
4) Holding Voltage (Vhold), this is the initial voltage that the SCR snaps back to.
5) Holding Current (Ihold), this is the minimum current required to keep the SCR in the low impedance region.
6) Thermal Breakdown (It2), this is maximum current that the SCR can tolerate before it gets damaged.

Therefore for ESD SCR design the safe region of operation is below BV or Vtrig and below Ihold. Hence the load operation of the pin must be inside this region as per Fig 32, over the full temperature range and max rating supplies.

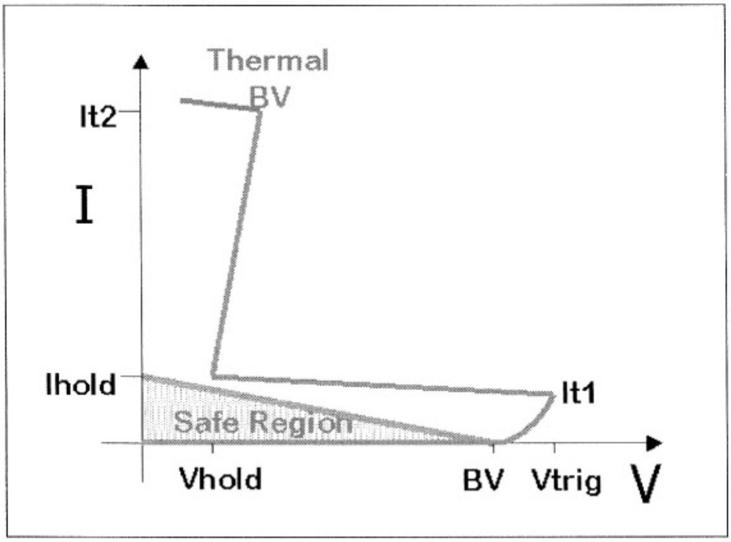


Fig.32. SCR IV Characteristic

Not exposing a pin to voltage glitches is very difficult to eliminate completely, therefore the Ihold requirement is the most critical, so that accidental triggering of the SCR cannot be sustained by the load line of the part and will therefore come back out of the low impedance region after the glitch has passed.

For SCR design the critical design parameters are;

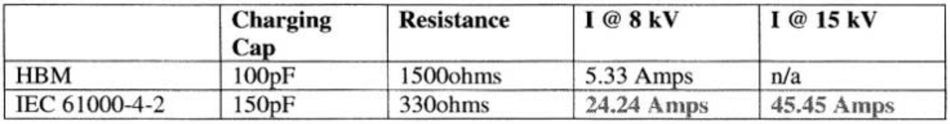
- Anode to Cathode spacing; affects the ESD rating, Vhold and Ihold
- Well Sheet Resistance; affects Ihold and Vhold
- Substrate Sheet Resistance; BV, Vtrig and It1

Unfortunately there is a reverse correlation between the Ihold & the ESD performance, therefore increasing Ihold for safe operation reduces the ESD performance of the SCR. So as always there is a trade-off between part performance and ESD performance.

10. 15 kV IEC ESD

International Electrotechnical Commission (IEC) ESD Standard 61000-4-2 is a System level ESD Standard. This system is adapted for IC I/O pins that can 'see' the outside world and therefore can be ESD stressed in their application. These would include IC pins tied directly to external connectors, Interface Line Driver/Receiver pins (e.g. RS232), Touchscreen pins, all of which may be exposed to external ESD pulses or Hot Socketing.

The key difference between this standard and HBM/MM is that the part may be powered up. Comparing the IEC model to the HBM model in Fig 32;



	Charging Cap	**Resistance**	**I @ 8 kV**	**I @ 15 kV**
HBM	100pF	1500ohms	5.33 Amps	n/a
IEC 61000-4-2	150pF	330ohms	24.24 Amps	45.45 Amps

Fig.32. IEC 61000-4-2 comparison to HBM Std

Therefore this is a far more severe standard than the HBM as the current levels for the same voltage is approximately 5x.

This ESD testing is carried out with an ESD Gun with interchangeable discharge tips to allow Contact or Air Discharge. Typical voltage rating requirements are 8 kV Contact & 15 kV Air Discharge. Fig 33 shows a part getting IEC zapped for 15 kV Air Discharge.

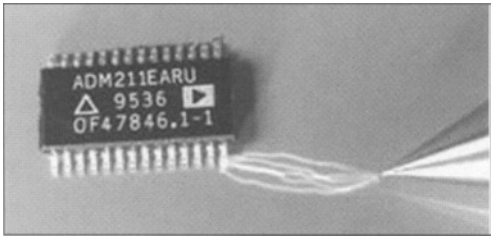


Fig.33. IEC 15 kV Air Discharge ESD zap

About the only ESD structure that is capable of handling this level of ESD current is a SCR, due to it's strong snapback mechanism and low Ron after snapback.

The standard SCRs are unipolar structures due to the parasitic diode to substrate in the negative direction. For Interface applications a bi-directional BV is required. This can be achieved with a floating Well structure, which has symmetrical Anode & Cathode regions and may include a Punchthrough diffusion to make the BV tunable [25] as per fig 34. The N+ isolation ring is required to prevent Latch-up to neighbouring active circuitry.

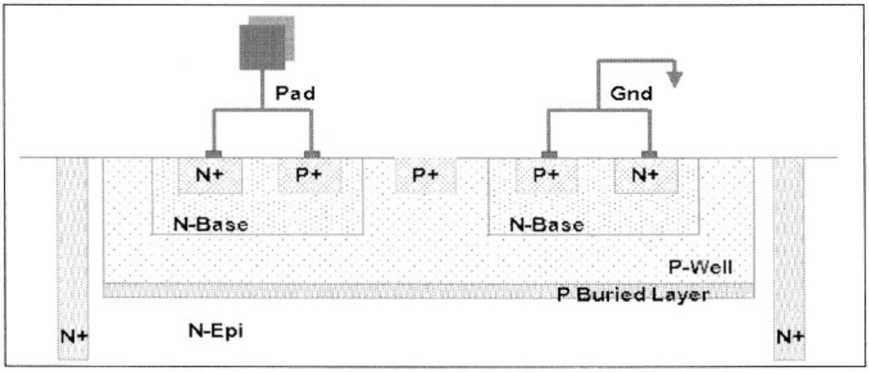


Fig.34. Floating Well Bi-directional BV SCR

In order to handle the high currents involved the structure needs a large active area, therefore an annular structure enclosing the Bondpad suits perfectly. Fig 35 shows an Photon Emission image of the annular nature of the 15 kV SCR structures with one of the SCRs at the point of snapback. Snapback initially occurs in the corners due to the increased Field Intensity in the curved corners.

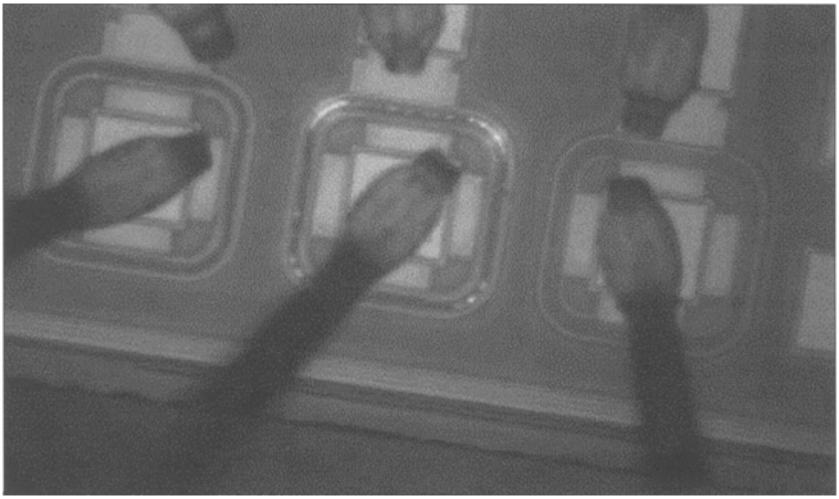


Fig.35. Photon Emission image of 15 kV SCR at snapback

Fig 36 shows the Photon Emission image of the SCR after snapback with the full SCR junction turning on.

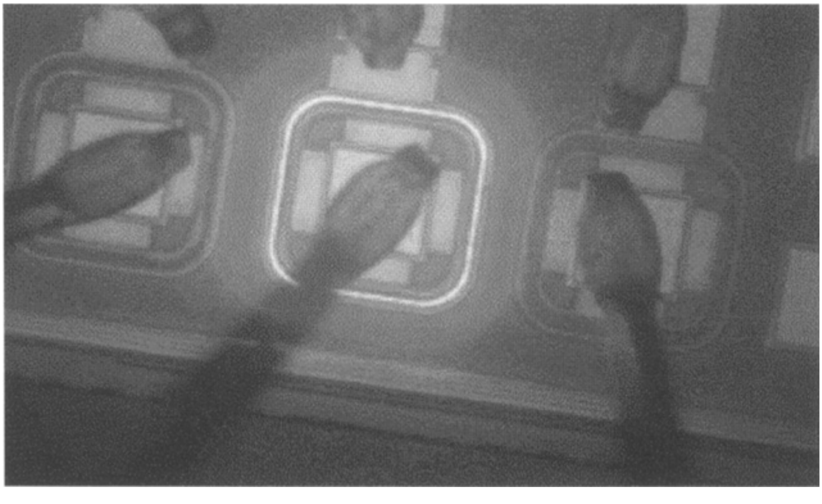


Fig.36. Photon Emission image of 15 kV SCR after snapback

11. BiCMOS ESD Failure Analysis Case Study

This is an example of the debug of an ESD weakness on silicon. The part is on a BiCMOS technology and the part failed 1.5 kV HBM on a Bipolar Input pin. Failure Analysis detected the fail site at the highlighted NPN Base Collector junction in Fig 37.

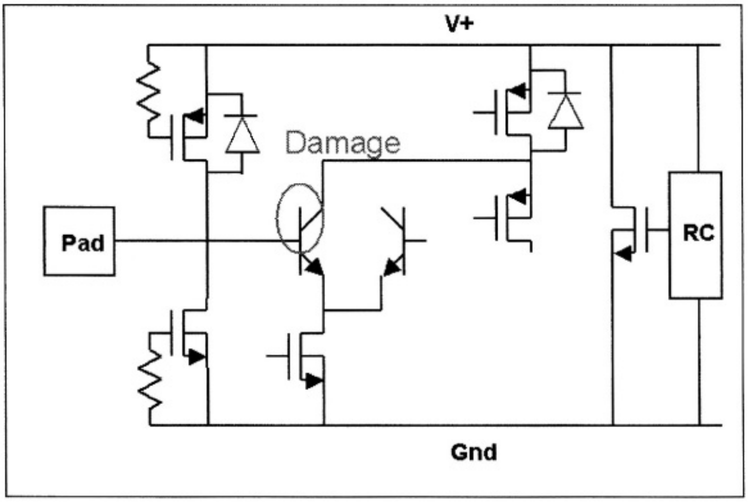


Fig.37. Schematic showing ESD damage site

With ESD zap experiments it was determined that damage occurred due to a positive zap from the Pad to Ground. The expected ESD path for this ESD zap is forward biasing the parasitic diode in the ESD PMOS to the floating V+ rail and snaping back the RC timer triggered Power Supply Protection to Ground as per Fig 38.

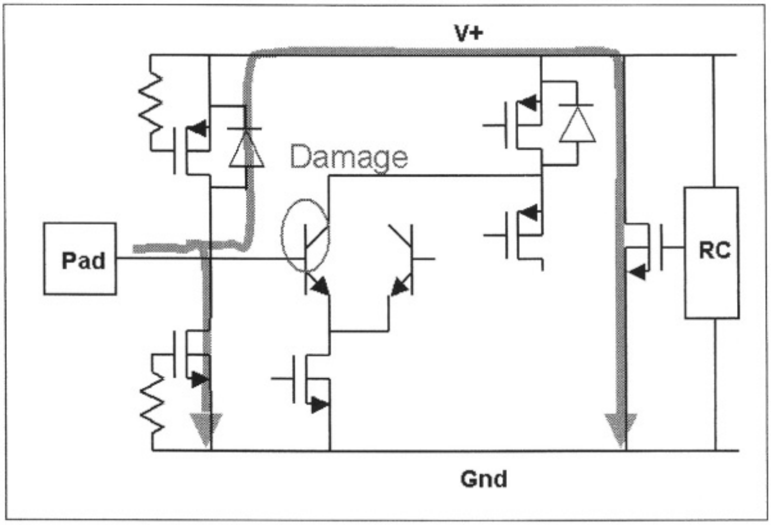


Fig.38. Expected ESD Current path for positive zap to Ground

Using the Focused Ion Beam (FIB), we were able to isolate the various ESD paths and with the TLP tester proved that 30% of the ESD current travelled through the NPN path as shown in Fig 39. Even though there was an extra diode drop in the path, the expected ESD path still took the majority of the ESD current but 30% of the ESD current at 1.5 kV was enough to damage the non-ESD NPN device. The fix was to include an ESD resistor to the Base of the NPN, to isolate the internal paths from the pad.

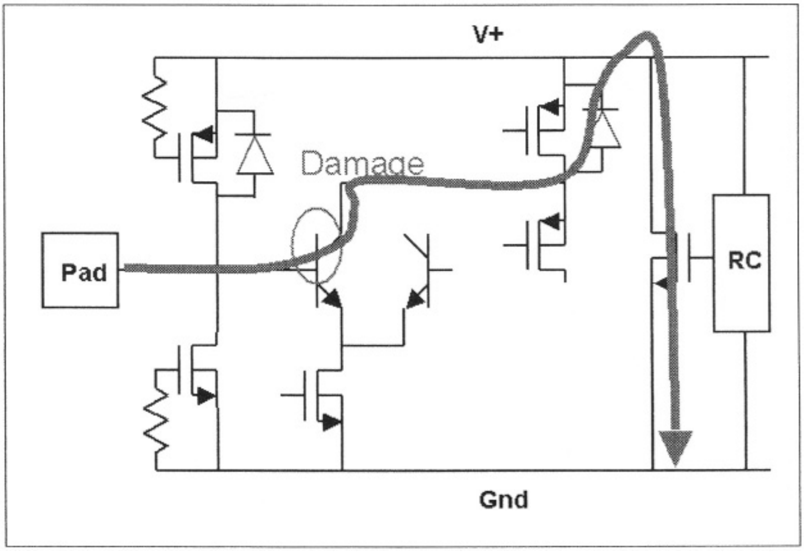


Fig.39. ESD current path causing failure

This ESD Case Study shows the importance of knowing the potential ESD paths, including parasitic diodes. And to understand that the 1st ESD path may not be the only ESD path, as the pad voltage increases due to Ron, then a 2nd or multiple ESD paths may kick in.

An automated checker is the ideal way to check this, whereby all the potential ESD paths are calculated based on the snapback and breakdown parameters of ESD and active devices and also including voltage drops caused by the impedances of the interconnect and rails.

12. Salicided Process Case Study

Deep submicron Salicided Processes use a low sheet resistance silicide layer on the Source/Drain diffusions. This silicide layer is weak for ESD since it is so shallow. For ESD devices which include ESD spacings from the Drain contacts to the Poly Gate, this spacing is shorted out by the silicide layer as per Fig 40 and will degrade the ESD performance dramatically. Therefore the solution is to use a Resistor Protector Oxide (RPO) block mask, which blocks out the silicide formation in the Drain region as per Fig 41.

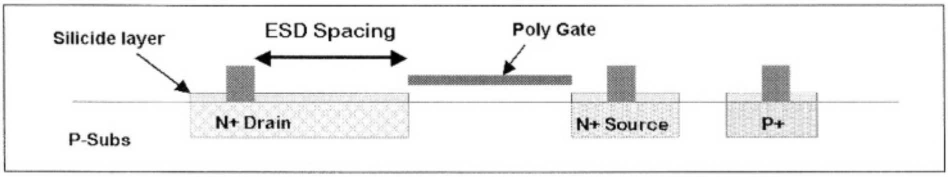


Fig.40. Salicided NMOS Cross-section without RPO

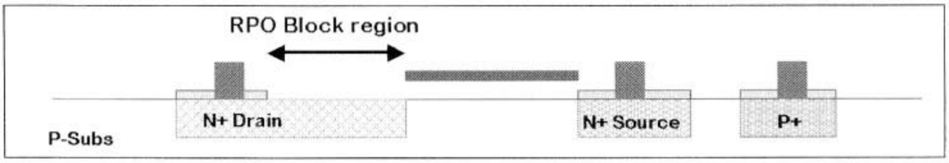


Fig.41. Salicided NMOS Cross-section with RPO

A part on a 0.25um Salicided Process had 500V HBM ESD failures. Failure Analysis pin-pointed the problem to a leakage to Ground in the NMOS ESD Cells as per liquid crystal image in Fig 42, hi-lighting the Hot Spot.

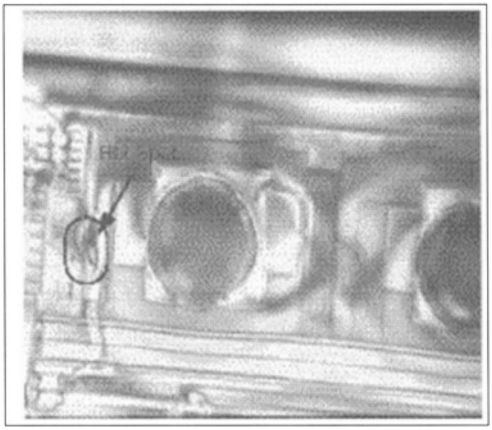


Fig.42. Liquid Crystal image showing Hot Spot

Using the Focused Ion Beam (FIB) a cross-section of the fail site was performed and the profile showed silicide at the Drain side of the Poly Gate as hi-lighted in Fig 43, indicated by the dips in the silicon surface.

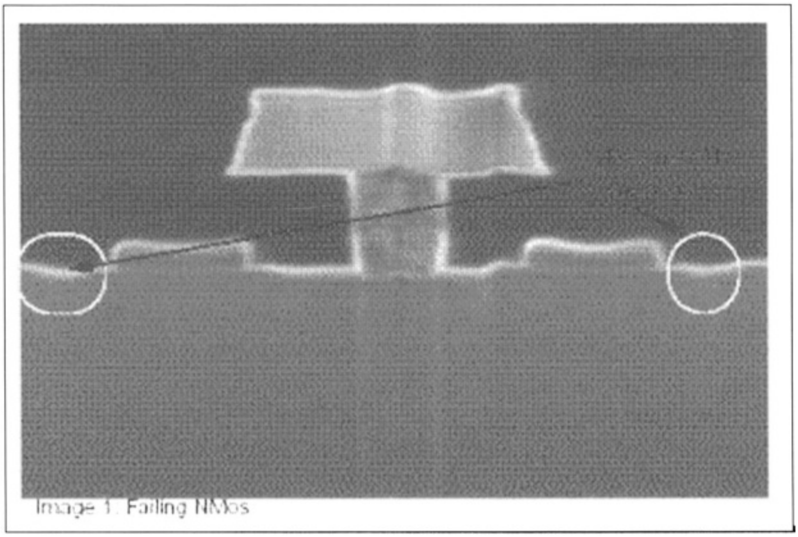


Fig.43. FIB Cross-section of ESD failure NMOS

The typical expected cross-sectional profile of an ESD NMOS is shown in Fig 44, whereby the silicide dips only occur on the inside of the 2 Poly Gates towards the Source contact (not visible).

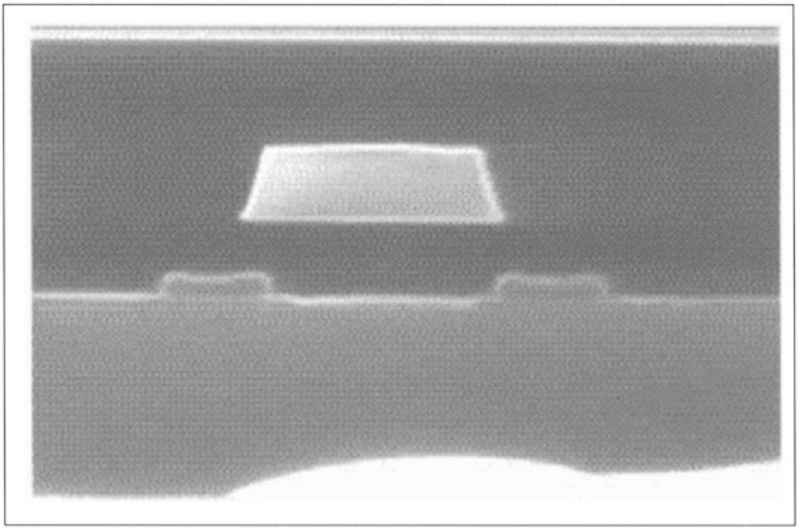


Fig.44. Typical Cross-section of NMOS device

A Transmission Electron Microscope (TEM) cross-section of the fail site showed the ESD damage at the Source contacts as per Fig 45, with the line of source contacts in the background.

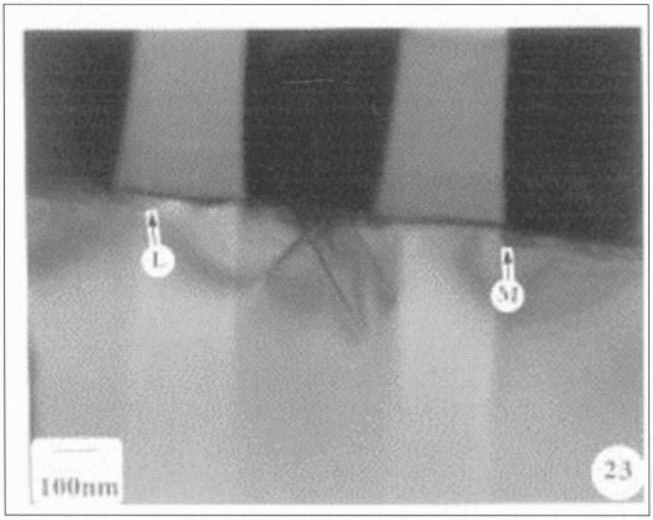


Fig.45. TEM of Fail site

This ESD weakness was caused by an error in the RPO Mask generation whereby a thin strip of silicide remained at the Drain side of the Gate as per Fig 46, which is a known source of defects for Salicided Processes.

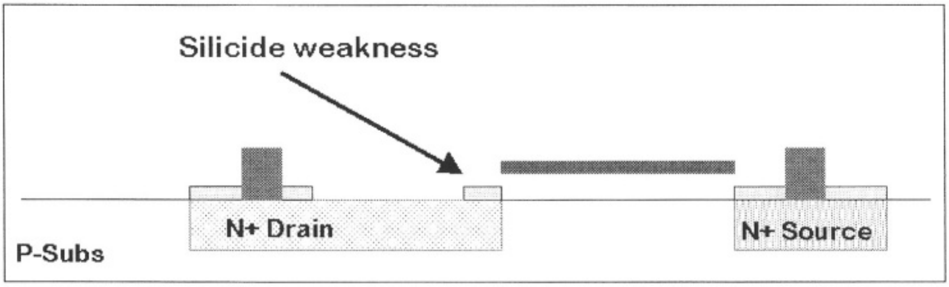


Fig.46 Error in RPO Generation Mask

This Salicide Case Study shows the importance of Failure Analysis in the debug of ESD issues on silicon, detecting the fail site and showing the type of damage, which gives critical clues to solving the ESD weakness.

13. ESD Design Techniques

There are some critical tools to help in ESD Design;

- TCAD: Process Simulations of ESD structures to understand operation, debug issues, improve performance & narrow down options of new ESD Cells for Testchip verification

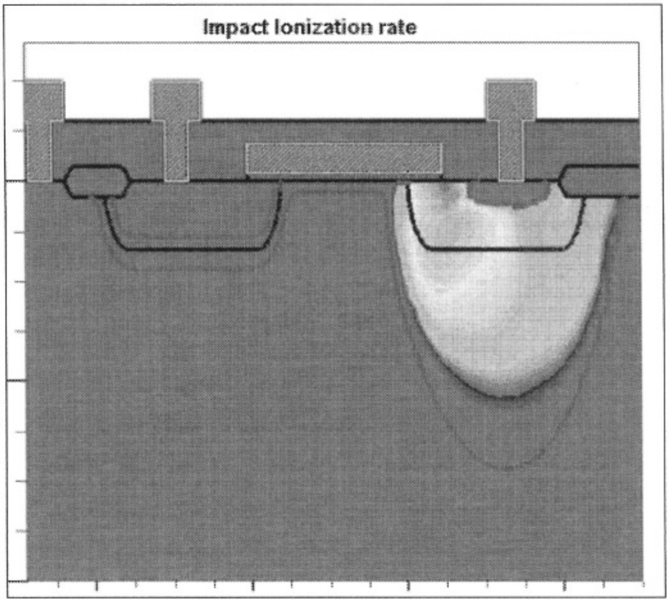

Impact Ionization rate

Fig.47. TCAD image of Impact Ionization at snapback

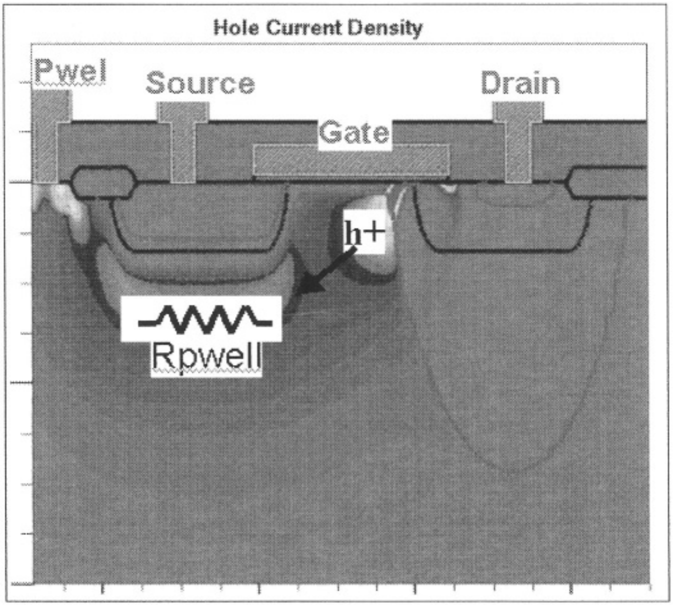

Hole Current Density
Pwel
Source
Drain
Gate
h+
Rpwell

Fig.48. TCAD image of Hole Current Density at snapback

- Testchip verification to develop new ESD Cells & Protection techniques, improve performance & generate ESD Design Rules

- Transmission Line Pulse (TLP) Testing to evaluate & characterise ESD Cells & debug issues on silicon
- ESD Schematic Map to hi-light all ESD paths (including parasitic paths) for every ESD pulse pin pair combination. This ensures that the ESD Protection is adequate and that all potential ESD paths are covered
- ESD Design Rule checker for layout verification prior to tapeout.
- Failure Analysis of testchip structures & silicon issues to debug ESD issues and feedback for generation of new ESD Design Rules.
- SPICE simulation where models include post BV & snapback characteristics. Parasitic Impedances & capacitances of interconnect needs to be also included, as these will cause voltage drops in ESD paths.
- Electrothermal Simulation, where the thermal effects of the ESD pulse can be simulated, which is especially important at the 2nd breakdown point (It2).

Issues to watch out for in ESD Design;

- Understand ESD Paths and make sure that all pin combinations are covered
- Take into account the Parasitic Diodes associated with MOS & Bipolar devices
- ESD spacings required on Source side of ESD NMOS if Source isolated from backgate
- Watch out for Parasitic devices in layout between I/O pads and/or Power Supplies/Grounds, in particular Lateral NPNs.
- Power Supply Rail and connection impedances as these cause voltage drops in ESD path
- Ensure that there are ESD paths to protect interface circuitry between blocks, especially those which operate at different voltage levels.

14. Conclusions

Analogue ESD Design is more difficult than Digital ESD Design due to the more extensive applications that Analogue pins may have e.g. HV, RF, Low leakage etc. Usually this requires the development of a new ESD Cell for the particular application.

TCAD, Testchip Evaluation and TLP characterisation can help dramatically in the development of these new cells prior to tapeout.

Be flexible with the design in allowing easy modifications in Metal e.g. to allow ESD resistor sizes to be tweaked or to metal in alternative ESD structures.

If ESD weaknesses do occur, then Failure Analysis is critical and when the weakness mechanism is understood, to include a new Rule in your ESD Design Rule Specification and Checker program to eliminate future occurances.

There are some excellent ESD Design Text Books available [26], [27], [28], which cover all areas of ESD Design from Device Physics right through to advanced Simulation.

References

[1] ESD-STM5.1-2001, "ESD Association Standard Test Method for Electrostatic Discharge Sensitivity Testing: Human Body Model (HBM) – Component Level", ESD Association, Rome, NY, 2001.

[2] ESD-STM5.2-1999, "ESD Association Standard Test Method for Electrostatic Discharge Sensitivity Testing: Machine Model (MM) – Component Level", ESD Association, Rome, NY, 1999.

[3] ESD-STM5.3.1-1999, "ESD Association Standard Test Method for Electrostatic Discharge Sensitivity Testing: Charge Device Model (CDM) – Component Level", ESD Association, Rome, NY, 1999.

[4] "Analog Devices EOS/ESD Reliability Handbook". 2000.

[5] D. Scott, G. Giles and J. Hall, "A lumped element model for simulation of ESD failures in silicided devices", 8th EOS/ESD Symposium, pp.41-47, 1986.

[6] K-H. Oh, C. Duvvury, C. Salling, K. Banerjee and R. Dutton, "Non-uniform Bipolar Conduction in Single Finger NMOS Devices in a 0.13um CMOS Technology", IRPS proc, pp.226-234, 2001.

[7] K. Chen, "Effect of interconnect, process and snapback voltages on the ESD failure threshold of NMOS transistors", 10th EOS/ESD Symposium, pp 212-219, 1988.

[8] T. Polgreen and A. Chatterjee, "Improving the ESD failure threshold of silicided nMOS output transistors by ensuring uniform current flow", 11th EOS/ESD Symposium, pp. 167-174, 1989.

[9] C. Duvvury and C. Diaz, "Dynamic Gate-Coupled NMOS for Efficient Output ESD Protection", 30th IRPS proc, pp.141-150, 1992.

[10] J.Z. Chen, A. Amerasekera and C. Duvvury, "Design Methodology for Optimizing Gate Driven ESD Protection Circuits in Submicron CMOS Processes", 19th EOS/ESD Symposium, pp.230-239, 1997.

[11] C. Duvvury, C. Diaz and T. Haddock, "Achieving Uniform nMOS Device Power Distribution for Sub-Micron ESD Reliability", IEDM Tech Dig, 1992.

[12] W.R. Anderson, "Circuit and Process Design Considerations for ESD Protection in Advanced CMOS Processes", Microelectronics Reliability

37, pp.1087-1103, 1997.

[13] D. Krakauer and K. Mistry, "Circuit Interactions during Electrostatic Discharge", 16th EOS/ESD Symposium, pp.113-119,1994.

[14] C. Richier, N. Maene, G. Mabboux and R. Bellens, "Study of the ESD Behaviour of Different Clamp Configurations in a 0.35um CMOS Technology", 19th EOS/ESD Symposium, pp240-245, 1997.

[15] L. Singer, "Method & apparatus for providing ESD/EOS protection for IC power supply pins", US Patent 5838146, 1998.

[16] H. Chang & M. Ker, "Improved Output ESD Protection by Dynamic Gate Floating Design", IEEE trans on Electron Devices, 1998.

[17] D. Beigel at al, "IC with 2 terminal Diode device to protect Metal-Oxide-Metal Capacitors from ESD Damage", US Patent 5477078, 1995.

[18] Oryx Instruments Celestron 1 TLP Tester

[19] Barth Electronics TLP Application Notes

[20] T. Moloney and N. Khurana, "Transmission Line Pulsing Techniques for Circuit Modeling of ESD Phenomena", 8th EOS/ESD Symposium, pp.49-54,1985.

[21] L. Avery, "Using SCRs as transient protection structures in integrated circuits", 5th EOS/ESD Symposium, pp. 177-180, 1983.

[22] R. Rountree, "A Process tolerant input protection circuit for advanced CMOS processes", 10th EOS/ESD Symposium, pp201-211, 1988.

[23] A. Chatterjee & T. Polgreen, "A low voltage triggering SCR for on-chip protection at output & input pads", Electron Letter Device, EDL-12, p 21-22, 1991.

[24] B. Carbajal, R. Cline and B. Anderson, "A Successful HBM ESD Protection Circuit for Micron & Sub-Micron Level CMOS", 14th EOS/ESD Symposium, pp.234-242, 1992.

[25] M. Daly, "SCR Cell for Electrical Overstress Protection of Electronic Circuits", US Patent 6236087, 2001.

[26] A. Amerasekera and C. Duvvury, "ESD in Silicon Integrated Circuits", John Wiley & Sons, 2002 2nd Ed.

[27] A. Wang, "On-Chip ESD Protection for Integrated Circuits, An IC Design Perspective", Kluwer Academic Publishers, 2002.

[28] S. Dabral and T. Moloney, "Basic ESD & I/O Design", John Wiley & Sons. 1998.

ESD IN SMART POWER PROCESSES

G. Croce[1], A. Andreini [2], L. Cerati[1], G. Meneghesso[3], L. Sponton[1]

[1] STMicroelectronics, TelecomPeripheralAutomotive Groups R&D,
Via C. Olivetti 2 Agrate Brianza, Milano,Italy
[2] STMicroelectronics, TelecomPeripheralAutomotive Groups R&D,
Via Tolomeo 1, Cornaredo, Milano,Italy
[3]Padova University, DEI,Via Gradenigo 6/A, 35131 Padova, and INFM-Padova, Italy

1. Introduction

Electrostatic Discharge (ESD) is today one of the major concern for Integrated Circuits (IC) chip quality. Amerasekera and Duvurry [1] have identified ESD as one of the most common failure mode in silicon ICs: roughly 10 % of the failing parts is attributed to ESD event. Vinson and Liou [2] reported that roughly 20 % of the field return failures may be attributed to ESD while Wagner *et al.* [3] identified ESD as being the cause of greater than 25% of the failures encountered. Chip redesign and complete new mask set are often required to solve ESD weakness detected in the product development phase. Costs associated with these values are huge, but even more difficult to measure are the intangible costs related to delays and to the loss of customer confidence caused by failed products. Advanced process technologies did have a severe impact on ESD robustness: linear dimensions (channel lengths, gate oxide thickness and junction depth) scaling in new processes comes together with supply voltage reduction but "unfortunately" people do not scale accordingly! For this reason the ESD level requested is not reduced in new technologies and many new applications, made possible by advanced IC processes, require even increased ESD performances. Furthermore advanced technologies allow designers to realize very complex system on chip (SOC) with very large pin counts. The bonding pad pitch has to scale properly and the total area available for ESD protection circuit is going to become smaller and smaller. Hence, the ESD protection efficiency ($V/\mu m^2$) has to increase in the new process platforms. In order to achieve this task it's very important to take advantage of the dimensions shrinking without loosing too much in protection linear performance ($V/\mu m$). New materials introduction may bring some benefits like for example contact barrier, but they can also worsen the ESD protection

A. van Roermund et al. (eds.), Analog Circuit Design, 169-206.
© 2003 *Kluwer Academic Publishers. Printed in the Netherlands.*

performance as in the salicide case [1]. At the same time advanced SMART Power technologies [4] [5] complexity is growing accordingly to new products requirements. The present trend [6] is to offer an unique process platform for very different products' fields: hard disk, printers, LCD drivers, RF power amplifier, Wireless power management, automotive, etc. As the reliability and quality requirements are normally driven by the application field, the ESD protection structures in Smart Power technologies have to be optimized for a very widespread specification range.

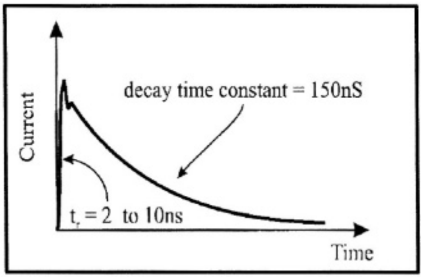
Normally on the same chip complex logic functions are put together with analog blocks and high voltage (from 5 to 70V is the most part, but several applications are covered up to 700V) output circuits. In order to meet the new products requirements, high voltage Lateral DMOS transistors are integrated in advanced CMOS processes [5][6]. Bipolar transistors are offered for particular analog functions even if the current trend is to minimize their use. Historical BCD (Bipolar CMOS DMOS) application field is the motor driver field where the DMOS transistor performances (high blocking voltage, BV_{dss}' low on state resistance, $r_{ds,on}$) make this device as the best choice for electronic power switches. However parasitic current injection in the substrate is the price that has to be paid when inductive load are driven: ESD protection structure has to be properly designed in order to avoid latch up issues triggered by the "difficult' IC environment. In this paper we will give an overview of ESD protection structure taking as a reference high density BCD processes and focalizing our attention on the high voltage pins protection strategy.

2. Stress Models

As reported above, ESD is a subset of a broad spectrum of electrical stress known as electrical overstress, EOS.

An ESD event occurs when a charge unbalance exists between an IC and another object. There are three major charge generation processes: triboelectrification (exchange of electrons between an insulator and another material in contact: residual charge remains on the insulator), induction (a charged object induces charge separation on another neutral object. Grounding the latter removes polarized charge so that when the charged object is removed a net charge remains on the initially neutral object), conduction (charged object in contact with a neutral one. An exchange of charge occurs, equilibrium is reached when the two objects reach the same potential: charge on the two objects depends on their capacitance).

Many efforts have been done in order to minimize charge build up or charge transfer between ICs and charged objects both in terms of facility automation



(reduced human handling) and in terms of new materials introduction in back end lines (assembly and final testing).

Nevertheless it's easy to imagine that it's pretty impossible to avoid the occurrence of an ESD event during the cycle undergone by the IC to become a finished good.

The IC industry has standardized on three basic models to define how the charge is transferred during the ESD event.

There are of course other models very peculiar to particular applications (IEC 1000, for example) but they will not be treated in this paper.

The standard models developed are known as Human Body Model (HBM), Machine Model (MM) and Charged Device Model (CDM): their names derive from the origin of the ESD pulses.

Brief details of the three models are given in this section.

HBM (Human Body Model) is the most popular ESD model and it's one of the most used methods to compare ESD threshold levels. It has been conceived in order to reproduce the ESD event generated by the discharge of a human being through a low impedance path and it is specified in the MIL-STD 883C method 3015.7.

Typical waveform is reported in Fig. 1: current pulse risetime is between 2 and 10 nsec, while decay time is about 150 nsec. The pulse is generated by the discharge of a 100 pF capacitor through a 1.5 kΩ resistor (Fig. 2) into the device under test (DUT in figure 2). Peak current value depends on the capacitor precharge value.

Parasitic elements, always present in any commercial equipment and associated to wires and test socket, are also reported in the figure and they may modify the current pulse shape and peak value.

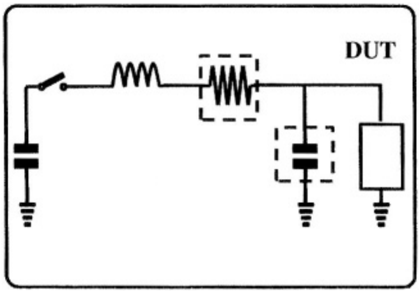


Figure 1: HBM current waveform

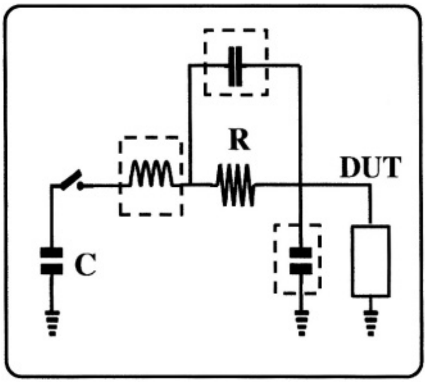


Figure 2: Circuit for ESD HBM and MM waveform generation. For the HBM: C=100 pF and R= 1.5 kΩ. Parasitic elements are reported inside dashed boxes.

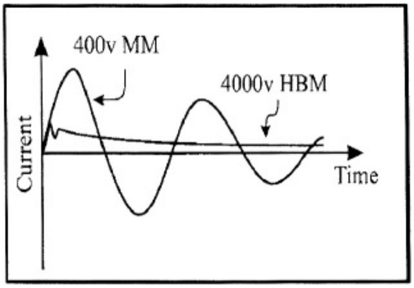


Figure 3: MM ESD with parasitic elements. Schematic and current waveform.

International standards exist establishing testing procedures (i.e. number of pulses, grouping of pins to test, etc.) and classifying circuits failures. For HBM 7 classes [7] are identified depending on the failure voltage level defined as the open circuit voltage of the charged capacitor. Normally failure voltage greater than 2 kV (Class 2) are mandatory: there is evidence,in fact, that products with ESD thresholds lower than 2 kV are prone to high failure rate during assembly and handling.

The Machine Model (MM) is very similar to the HBM. It was originated in Japan and it tries to model an ESD event induced by an automatic equipment.

The discharge capacitor in MM model is increased to 200 pF while the resistance value should be ideally 0. The inductance is attributed to the socket and package leads: typical value is 0.75 µH. Due to the resistance absence current peak values are much higher than HBM at a given precharge voltage: normally minimum MM voltage level required is 200 V.

The Charged Device Model (CDM) is the newest model and also the most difficult to reproduce the existing international standards are still being improved by therelevant commitees, mainly as far as calibration procedures and accuracy of metrology chains are concerned, to the aim of getting better repeatibility and migrability of the experimental results among different test equipments.

This model has been conceived in order to reproduce the event occurring from a charged package when one of the pin is grounded. CDM pulses have very fast duration, less than 1 nsec, and current levels may reach several amperes for a single event.

There are two methods for CDM testing that depend on how the part is held during the ESD testing: socketed and non socketed [8].

CDM event is very rapid (see Figure 4) and one of the main concern is the ESD protection turn on time which may be longer.

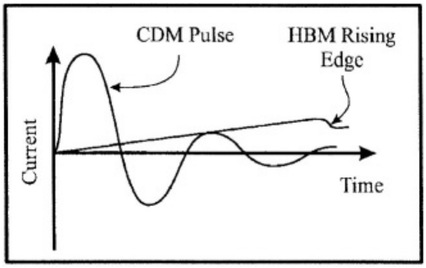


Figure 4: CDM current waveform.

In non socketed CDM ESD testing, the IC is charged to a precharge value and suddenly a pin is connected to ground.

CDM testing is coming recently a "hot" topic. In fact typical failures induced by this kind of testing are pretty different from the HBM and MM ones. Due to the large current involved and the very fast event, over voltage are generated that may induce oxide breakdown in the input stages due to reduced oxide thickness in modern technologies.

However since testing results are product/package dependent (the charge stored in non-socketed CDM testing depends on the "IC system" capacitance value) in this paper we will focalize on HBM results of elementary structures/cells.

Finally it's worthwhile to mention here a testing method today commonly used for ESD structure characterization. In order to model properly ESD protection structure it's important to study and characterize electronic devices' behaviour in the nsec time range. To this purpose Trasmission Line Pulse (TLP) testing is today commonly used [9].

Basically the system generates high value square current pulses (4-8 A) with 1 nsec rise time and 100 nsec pulse width in order to investigate circuits behaviour in the ESD time regime.

Today Very Fast TLP systems [10] have been presented capable of 250 psec rise time and pulse widths as short as 2 nesc in order to model protection behaviour for the CDM ESD time domain.

3. ESD protection network

In this section the different on-chip ESD protection strategies will be presented together with a generic description of what an ESD structure is expected to do. Basically each pad of an IC has to be protected against ESD event. Today two are the most common architectures for realizing an ESD protection network (Figure 5):

1) **In the Pad:** each pad is protected with an unique element put between the pad and a common line (normally the ground line).
2) **Common rail or Diode based:** each pad is connected through two diodes to two common lines (normally supply and ground lines). Between them an ESD protection structure is introduced.

In Figure 5 the two possible solutions have been reproduced in a simplified ESD network: both solutions in fact may be used in the same chip depending on the different pads specifications.

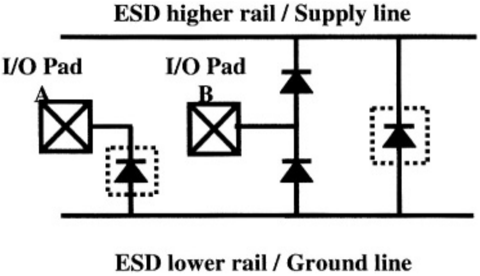


Figure 5: Simplified ESD protection network. Two I/O pads are protected with two different architectures. Structures in the dashed box are not simple diodes but real ESD protection structures.

The different solutions proposed, in fact, are not always interchangeable in the sense that specific pins specification may impose the use of one of the two.

Before going in further details a brief and simplified explanation of the working principle of the two architectures will be given.

The ESD protection structure put between the pad and ground in fig. 5 has to be

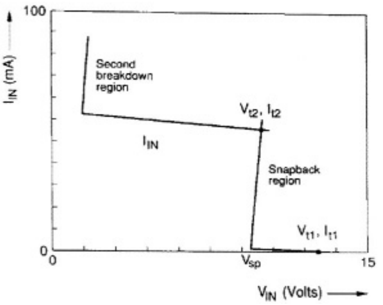


Figure 6: Generic I-V characteristic of ESD protection

designed in order to shunt the ESD current when a positive ESD pulse is applied. In figure 6 a typical I-V characteristic of protection structure is reported. Basically the device is triggered when a certain voltage is reached (trigger voltage, V_{t1} in fig.6). The pad voltage suddenly decreases to a lower voltage, usually indicated as “holding” or “snapback” voltage (V_{sp} in fig. 6) and the device enters in the snapback region. This is the region where the device has to work during the ESD event.

The voltage increases again when the current is increasing: the electrical resistance of the protection in this region has to be as low as possible. In fact typical failure of this kind of protection is thermal (or second) breakdown [1](V_{t2} in figure 6). Once the protection reaches this condition, it suddenly fails.

Conditions for second breakdown are defined by the internal temperature of the device: due to the power dissipation the internal temperature starts to increase up to limit value. This limit seems to be the intrisic silicon temperature, T_i, defined as the temperature at which the intrinsic carrier concentration in silicon, n_i, reaches the background doping concentration. Once this limit is reached silicon resistivity coefficient becomes negative and system enters in a positive loop.

A current increase in fact causes a resistance decrease so leading to a new current increase: temperature in the device may reach suddenly a critical value such as silicon melting limit.

Since the mechanism is thermally activated, for robust protection it’s mandatory to keep the temperature in the structure as low as possible. This can be achieved by minimizing the power dissipation in the device.

Power dissipation can be minimized by:

1) **keeping holding voltage as low as possible:** this can be achieved by a proper protection design
2) **minimizing the ESD protection electrical resistance in the snapback region:** this parameter will of course depend on the protection dimension. Proper protection design is anyway mandatory in order to reduce the specific resistance (resistance per unit area) and to ensure an uniform turn on of the structure.

Both items will be covered in next sessions where protection structure characteristics will be analyzed in detail.

When negative pulses are applied, the ESD protection structure is a forward biased diode. The voltage drop across it is pretty low and it’s limited by the diode series resistance: this means that the power is dissipated over entire device region so leading to larger ESD robustness.

Hence when “in the pad” architecture is used, the ESD current path between each couple of pads includes a structure working in the snap back region and the same structure working as a forward biased diode (unique exception is the discharge vs. ground pad).

The main limitation of this architecture is that each pad requires a dedicated protection structure that has to be sized in order to dissipate the ESD pulse working in the snapback region. Beside the overall area consumption, one of the major drawback is the capacitive load associated with the protection that may not he compatible with some pad specification especially for RF design [11].

Common rail architecture solves the capacitive load issue introducing a couple of diode at each pad. The diodes work in forward mode during the ESD event forcing the current through the ESD structure put between the high voltage(supply) and low voltage (ground) rails.

Theoretically, assuming zero the bus resistance, an unique ESD device has to be put for all the pads present in the IC.

This device may be again a protection structure identical to the one reported above, working in snapback region, or it can be a dynamically activated protection structure [12] [13].

Latter structure are today widely used: they are normally in off state and they are triggered by large dV/dt pulse typical of ESD event. For this reason (dynamic turn on) they are normally put between two lines with pretty stable voltage (supply and ground) in order to avoid undesirable turn on during normal IC operations.

Main advantage of this configuration is that ideally an unique device has to be properly sized (the clamp between the lines) while all the other components in the ESD current path are forward biased diodes that can be kept smaller due to their larger robustness. Pad capacitive loading is reduced as long as small area (perimeter) diodes are used.

Major drawbacks of the common rail solutions are:

1) Practically metal bus connecting the pad to the clamp have a finite resistance. For this reason the pad voltage during ESD event will increase due to the metal resistance ohmic drop. Voltage may exceed critical value defined by the circuit tied to the pad so leading to circuit failure. Therefore particular care must be taken on the placement of an adequate number of clamps.
2) Normally the two ESD rails coincide with the supply (V_{dd}) and ground (V_{ss}) lines in order to reduce ESD area consumption. If dedicated I/O supply pads are available (as in actual advanced Low voltage IC products) the overall Vdd line capacitance, due to the tied circuitry, is low and dynamic clamp may be used without problems. But if the line capacitance is high (as for

example in the core supply case) dynamic clamp may not be triggered by the ESD pulse. Duvurry [14] suggests that for supply line with capacitance larger than 10 nF protection device may be avoided.

3) Hot pluggable (or Failsafe [15]) I/O cannot be protected with this architecture unless dedicated floating ESD rails are used. In fact, in this case, pad voltage may exceed V_{dd} so turning the diode on during normal IC operation.

Another disadvantage of the common rail architecture is that two metal lines, that have to be sized properly in order to carry the ESD current and to minimize ohmic drop, have to be routed around I/O ring. Hence even if the single protection silicon area (diodes couple) can get smaller the overall protection network may require larger area. Routing constraints in the common rail architecture may be a limitation for the integration of ESD protection structures under bonding pads.

The final protection network choice cannot be performed without considering what is tied to the pad in Figure 5.

ESD protection structures robustness is in fact the first milestone to be reached, but it's only the starting point. The next step is to ensure that the identified structures are capable of protect the different circuit blocks that may be tied to the I/O pads.

Basically it has to be verified that pad voltage reached during an ESD event is "low" enough to guarantee the full functionality of the circuitry tied to it after the event.

This is a first condition that has to be satisfied otherwise even if the protection itself may safely work after an ESD event, circuits will fail. Failure may be catastrophic(oxide rupture or metal filament for example) and its electrical signature is normally a leakage increase [1].

In some cases, however, the failure may be parametric (ex. voltage shift) and may be revealed only after a complete IC testing or, and this is by sure the most dangerous situation since it can impact reliability of the ciruitry [16].

This consideration will impose the maximum voltage that a pad could reach during the event and it will of course depend on the pad circuitry (input,output) and the voltage class.

At the same time the protection structure must not intefere with IC normal operations: it must be off in the normal circuit operating region. Latter consideration fixes lower accettable limit for ESD protection trigger voltage.

In case of structure working in snapback region another key parameter is the holding voltage. As stated above in order to maximize protection robustness (i.e. minimize power dissipation), holding voltage should be kept as low as possible but this can cause some problems during normal ICs operation.

In fact accidental turn on of the ESD protection structure may occur. This can be caused by noise pulse that can dynamically trigger the protection: trigger voltage in fact may decrease when large dV/dt signal are applied to the protection structure [17].

In case of Smart Power ICs the noise level in the substrate is worsened by the current injected by the power LDMOS when inductive load arc driven. Parasitic currents in the substrate may trigger the ESD protection structure.

Once the protection is triggered, if the holding voltage is lower than the pad voltage the ESD structure would be kept in this state and it would sink current from the pad for a theoretically indefinite time. Pratically the protection cannot withstand the power dissipation for long time and it will blow out.

One way to solve this issue is, of course, minimizing noise propagation by using proper design tricks and, in case of Smart power process, by reducing current injection in the substrate and by insulating the protection structure from the substrate utilizing N buried layer.

This approach may have a major drawback: it's based pretty often on designers experience and will be product dependent.

A way to completely solve this potential issue is to keep the holding voltage greater than the supply one. So doing an accidental trigger of the protection would not be sustained by the pad voltage and the protection will immediately turn off. This solution has an obvious drawback: ESD protection performance could be worsened and larger area may be needed for same ESD protection level.

But the extra area is surely well compensated by the overall chip quality expecially in the very harsh environment typical of Smart power applications.

An example of a robust protection I-V characteristic is reported in Figure 7.

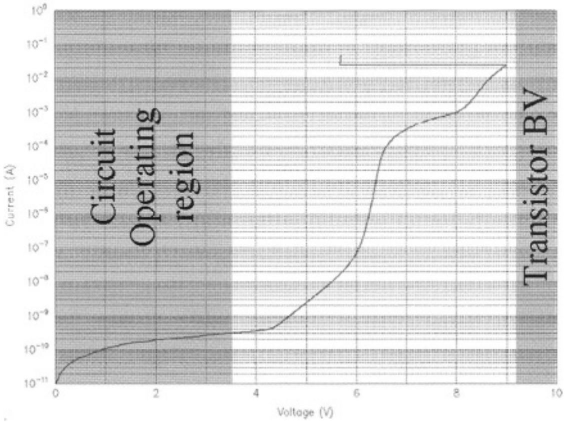


Figure 7: Quasi static IV characteristic of a 3.3 V ESD protection. Supply voltage and transistor breakdown are the evidenced regions.

The ESD structure will be described in detail in the next session: however from the analysis of the low current region reported in figure 7, trigger and holding voltage are pretty clear. As evident these values stay in what is normally called *ESD window* [18] : trigger voltage (V_{t1} φ 9 V) is below transistor breakdown value, that is the critical value for output pads, and holding voltage (V_h φ 6 V) is greater than maximum supply voltage (3.6 V for a 0.35 μm technology).

The ESD window so defined imposes several constraints on the protection structure design, particularly for the high voltage pins.

While in advanced CMOS process supply voltage scales accordingly with dimension reduction, this does not happen in Smart power process for the high voltage classes.

In this case, in fact, the supply voltage is imposed by the specific application and it does not always change in new products' generations: in some fields a supply voltage increase has been even introduced.

In automotive market, in fact, a battery voltage increase, from 14 to 42 V, has been proposed [19]. Such change will move the absolute maximum voltage to 58 V and when external drivers have to be driven (+12 V) this value is going to become 70 V [20].

Since normally 80 V power devices are used, the ESD window will be roughly 10 V: this means roughly 15 % of the full voltage range.

In the case reported in figure 7, typical for a 0.35μm LV CMOS process, the window is roughly 6 V for 3.6 V voltage classe: quite 200 % of the full voltage range!!

4. ESD protection structures

Within this session, an overview of the elementary protection structures used in smart power applications will be provided together with some typical stages design, basically input and output ones.

Both protection architectures proposed in the previous session will be considered and analyzed mainly for BCD6 [4] process platform.

This session will be divided into two different parts: the first one will be dedicated to LV logic circuit blocks while the second one will be focused on HV power blocks.

4.1. Low Voltage Circuitry

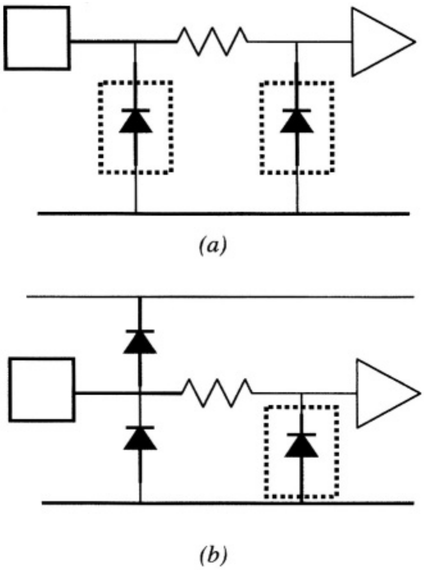


Figure 8: Conventional input pad protection scheme: (a) in the pad , (b) common rail network. The structure in the dashed box are not simple diode but ESD protection structure

Input stages

In the figure 8 conventional and simplified input pad architecture is reported [1].

Both “in the pad” configuration (8a) and common rail one (8b) are schematically reproduced.

In the first case two real ESD protection structure are necessary (dashed box).

The structure directly tied to the pad, *primary element,* has to shunt most of the ESD current while the structure tied to the internal node, *secondary element,* serves to limit the voltage at the circuit to be protected. A resistor is present in order to isolate the two elements.

The real need for the secondary element will depend on the efficiency of the primary one. Slow turn on or large trigger voltage, in fact, may cause a sudden increase of the pad voltage that may not be compatible with the circuit to be protected: it may exceed gate oxide breakdown voltage so leading to circuit failure.

However improvements in the ESD protection design may make the need of the secondary element useless. Existing margin between ESD trigger voltage and oxide breakdown, both under ESD pulse condition, has to be carefully evaluated in order to establish the real need of the secondary element.

However even if this element is not introduced, the resistor is normally left before the input stage so that it would limit the potential current flow through the oxide.

Today trends in advanced CMOS where the gate oxide thickness is scaling very aggressively is to keep the dual protection approach. This need may be forced by the recent interest in CDM technique [21]: in this case, in fact, fast pulse may not be immediately shunted by the primary device and a secondary one close to the gate will be very useful to avoid failure induced by the oxide breakdown.

Most widely used ESD protection structure for this architecture are: self triggered Lateral NPN (LNPN) bipolar transistors [1] and Lateral Silicon Controlled Rectifiers (LSCR) [1].

Bipolar lateral action is the basic mechanism for other structures like Gate grounded and Gate coupled MOS transistor that are actually the most used ESD structures in advanced CMOS processes. Even if they are realized like NMOS transistor, basic operation principle under ESD pulse is based on the intrinsic NPN transistor turn on: for this reason they will treated in the same session.

Before starting with the description of LNPN based structure, LSCR structure will be briefly analyzed.

In figure 9 Lateral SCR elementary section and equivalent schematic are reported.

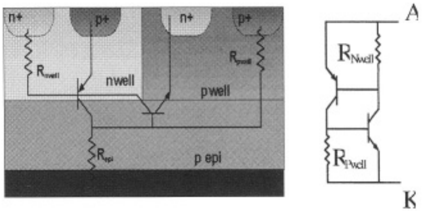


Figure 9: Lateral SCR cross section and equivalent circuit. Anode (A) terminal is tied to the pad to be protected.

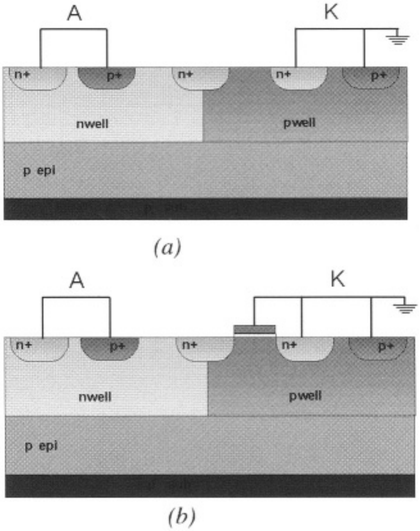


Figure 10: (a) Modified Lateral SCR (MLSCR) (b) Low-Voltage Trigger SCR (LVTSCR) schematic cross section.

The structure reported is the simplest that can be realized in a common CMOS platform where no buried layer is usually available.

In smart power technologies the same structure may be isolated by the substrate by using n buried layer or SOI (Silicon On Insulator) isolation scheme.

SCR protection structures are the most efficient in terms of ESD performance per unit area due to the very low holding value achievable.

Basic operation will be briefly described: the pad (anode) voltage increase during the positive ESD pulse will cause NPN base-collector breakdown. Avalanche current has to flowthrough Nwell resistance and it biases the PNP bipolar on. PNP collector current biases NPN transistor on so leading to the typical regenerative mechanism of latchup phenomena. Once the bipolars are switched on the holding voltage is very low:

$$V_h = V_{ce,pnp} + V_{be,npn} \varphi\ 1\ \mathrm{V}.$$

Hence power dissipation is very low and this is the reason for the excellent ESD performance: HBM robustness close to 90 V/µm has been recently reported [22].
In the structure reported in figure 9, trigger voltage is fixed by the NPN base (pwell) – collector (nwell) junction breakdown. The breakdown voltage value is normally pretty large (more than 20 V) and may lead to input stages fail.
Different tricks have been identified in order to reduce the trigger voltage (figure 10).
Basically a n^+ tap is added near the surface at the nwell edge: NPN base-collector breakdown value is reduced to the n+/pwell diode value (10 V range) and can be further reduced by adding a gate on this new junction (gated diode).
Major drawbacks of the SCR devices family are:

1) the turn on time is pretty high (compared to other protection structures) due to the presence of a p-type transistor to be turned on. This may limit severely CDM performances [23].
2) the very low holding voltage (1 V) makes the device unsuitable for smart power application for the consideration reported above. Moreover the need for low trigger voltage, due to the reduced oxide thickness in advanced CMOS platform, did reduce the noise margin immunity in CMOS process too [23]. Holding voltage may be increased but it will decrease ESD robustness.

Last consideration on ESD window has made SCR protection not widely used in smart power application.
Self triggered LNPN based protections are the most commonly used structures for low voltage pads protection in smart power processes [24] (figure 11).

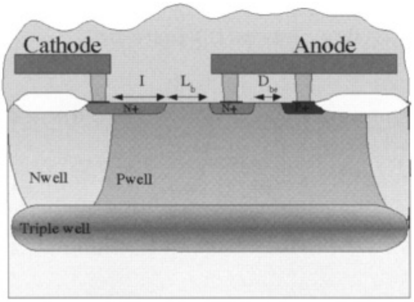


Figure 11: LNPN schematic cross section.

In figure 11 a schematic cross section of a lateral NPN feasible in any smart power technology is shown.

Some geometrical parameters are indicated since they play a meaningful role in the protection structure behavior and robustness:

1) I: it fixes the lateral NPN collector series resistance (n^+) value. Larger values do help in the structure uniform triggering but worsen ESD robustness because they lead to larger power dissipation.
2) Base width, L_b fixes the lateral bipolar gain and the structure speed. Structure with larger gain has lower holding voltage. Minimum value for this parameter depends on voltage capability.
3) Emitter-base distance, D_{be}, fixes the trigger current value since intrinsic base resistance is strongly influenced by this parameter.

Self triggered LNPN structure is basically a bipolar transistor with base and emitter shorted by a metal connection. During an ESD event the cathode (collector) voltage is increased up to the collector-base diode breakdown value. Electrons generated by the avalanche phenomenon are collected by the cathode while holes are collected by the base contact.

Holes' current induced ohmic drop through the intrinsic base resistance starts to forward bias the emitter-base junction so turning the transistor on. Once the transistor has been turned on, cathode voltage suddenly decreases (snap-back). Collector electric field necessary to sustain the external forced current may be lower since a larger amount of primary electrons is available when the bipolar is on.

In figure 12 dynamic I-V characteristics, obtained by a TLP measurement, of a real device realized in BCD6 process platform is reported.

Snap back (or trigger) and holding voltages are indicated in the figure together with the second (thermal) breakdown value.

As it has been said above, second breakdown is the mechanism responsible for the ESD protection failure: corresponding normalized TLP current level (often referred as I_{t2}) is the merit figure of the protection [25].

Fixed HBM failure threshold voltage and known the protection I_{t2} value, ESD structure dimension may be set.

Large I_{t2} values are obtained by minimizing power dissipation in the structure. This can be achieved by reducing the holding voltage as much as possible and by minimizing the protection equivalent resistance in the snapback region.

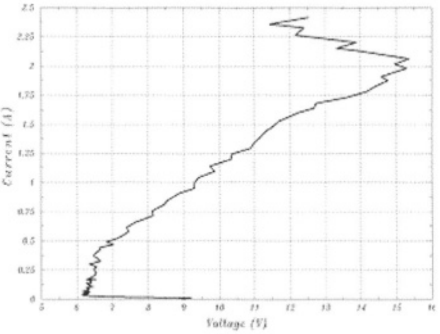


Figure 12: BCD6 LNPN TLP I-V characteristic.

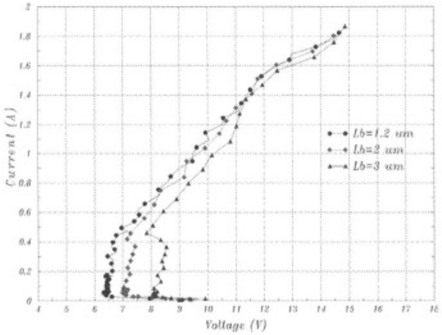


Figure 13: TLP I-V characteristic for structures with different Lb values (base length)

As evident from the figure 12 a sudden variation (increase) of the structure resistance occurs in the high current region (I>500 mA). In fact together with the designed lateral NPN transistor a parasitic one is present due to the presence of a n buried layer necessary for the structure isolation.

Current partitioning between the two transistors will depend on the geometrical factors above reported, base doping level and on the current level [26][27].

In figure 13 I-V characteristic for devices with different base width (Lb) values is reported. As expected longer bases lead to higher holding voltages, due to the lower gain. Hence in the low current regime lateral bipolar action is effective, while in the high current regime no main difference between the devices is observed due to the prevalent vertical action.

Vertical action may be helpful in improving ESD robustness since power is dissipated through a larger volume but at the same time particular care has to be taken in order to keep vertical electrical resistance as low as possible. Sinker implant presence and high doped buried layer are very helpful, when present, in achieving this task [28].

LDMOS transistors body implant (Pbody) may be successfully used as base implant for improving structure ruggedness [24] [29] HBM failure level in the 40-60 V/μm has been achieved with self triggered LNPN realized using Pbody implant. The result reported by Depetro et al. [29], HBM failure level φ60 V/μm with PBODY LNPN, has to be underlined since obtained on an SOI wafer with a pure lateral structure (no buried layer) and with the well known heat dissipation issue associated with this kind of substrate.

This result does confirm that lateral structures may be very robust if properly designed.

However unfortunately Pbody implant dose increase in advanced BCD platforms makes this implant unusable for the low voltage CMOS class since the base collector diode enters in the zener regime under operating condition (high leakage from the pad). Different structures coming from standard advanced CMOS processes are becoming very popular in smart power platforms.

Gate grounded devices are the most diffused protection structures in pure CMOS technologies: they are basically standard NMOS transistors whose gate is tied to ground (figure 14).

When possible these devices have no LDD implants as they decrease ESD robustness [1].

Working principle of the GGNMOS is the same as bipolar one: ESD event triggers the intrinsic parasitic bipolar which enters into snapback regime. These devices have some advantages:

1) They don't require any additional mask on a CMOS process
2) They have a very low triggering time: bipolar base length is defined by the gate length, which is usually the minimum lithographic dimension available

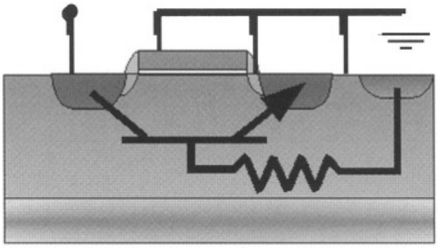


Figure 14: Gate grounded device section

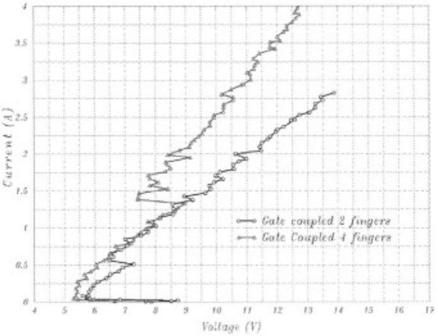


Figure 15: Gate coupled TLP I-V characteristics

It is common to increase drain (collector) resistance by keeping the contacts far from the gate edge in order to increase ESD robustness. Salicide technique, commonly adopted in advanced processes, will significantly reduce the effect and to reach good ESD levels it should be necessary to find other space consuming solutions (poly cut, stacked vias [30]) or to add a salicide protection mask that is however always available in smart power platforms.

The major drawback of the GGNMOS, like all the structures working in snapback region during the ESD event, is the current uniformity along the structure. In fact normally crystalline impurities or defects are present along the structure: breakdown does normally occur first locally in these points and then it extends along the whole structure [31]. If large structures with several fingers are designed it may happen that part of the structure enters in the snapback region before the rest of the structure (fig. 15). When this happen, if the second breakdown voltage (V_{t2}) is lower than the trigger voltage the device fail before the rest of structure triggering. GGNMOS has normally very high trigger voltage and they may suffer of poor robustness scalability.

This is the basic reason for the introduction of the Gate Coupled device (GCMOS, figure 16): the working principle of this structure is exactly the same of GGMOS but gate is biased both capacitively during the ESD event and statically during "slow" pulse through the diode. So doing larger current flows through the structure (standard MOSFET conduction current), hence lower electrical field are necessary for the bipolar base bias current generation. Smaller trigger voltage, V_{t1} is achieved together with uniform triggering along the finger (figure 17 and figure 18).

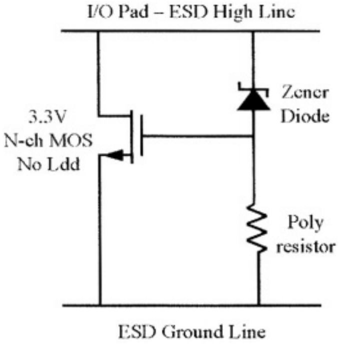


Figure 16: Gate coupled MOS schematic

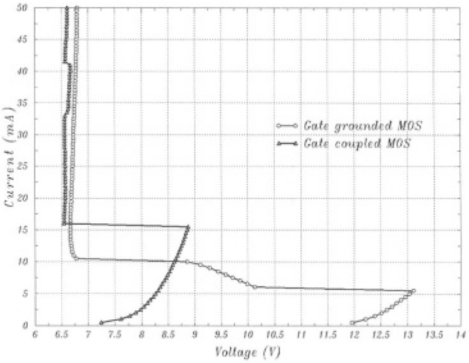
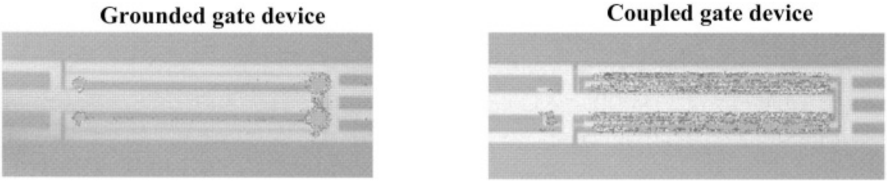


Figure 17: Gate Grounded and Gate coupled quasi static characteristics comparison: gate coupled device has a lower V_{t1}



Grounded gate device

Coupled gate device

Figure 18:Emission Light Microscope (EMMI) photos of a grounded gate (a) and a coupled gate (b) device , obtained stressing the structure with current pulses on drain side. Uniform triggering is achieved only in the gate coupled structure

In latest BCD technologies GCMOS devices are widely used to protect 3.3V circuitry, as they show a good robustness level of 25V/µm, and a good scalability with device width which has been proven till 12kV (figure 19).

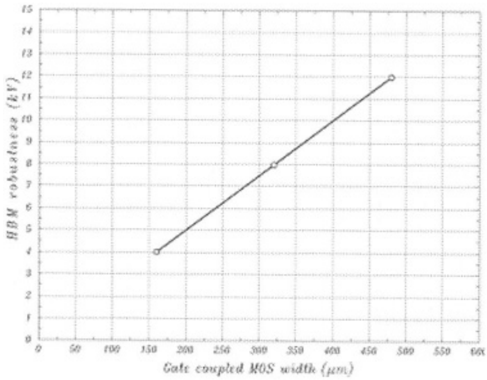


Figure 19: GCMOS HBM robustness

Recently Duvurry [32] has shown a novel tecnique for gate grounded robustness improvement: "substrate pump". Basically the base of the parasitic bipolar (normally coincident with the wafer substrate in a CMOS process) is directly biased by a transistor switched on during the ESD event. So doing no high electric field are basically necessary for the base current generation and the trigger voltage is hence reduced.

From the above reported HBM failure voltage level for bipolar-based protection structure it's easy to imagine that protection width in the 100-200 λm range has to be used for safely protect each pad (assuming 2 kV HBM requested level). Since normally large ballast resistance are introduced for robustness reason it's easy to image that capacitive load (both perimeter and area contribute) may become too large for peculiar application (value in the 1 pF range has to be considered).

Common rail architecture may offer a valid alternative when low capacitive load is mandatory as in RF designs.

Diodes used for building the common rail network are standard CMOS ones: Nplus/Pwell and Pplus/Nwell diodes opportunely sized to handle the large current during ESD event. Such diodes are really weak in reverse mode (robustness <1 kV) and present intrinsic parasitic bipolars (figure 20). Particular attention must be given in designing common rail networks to the non-idealities of these diodes:

1) Diodes have a finite on resistance that is not negligible and that contributes to the total impedance seen from pad to ground through the protection.
2) Parasitic elements of the diodes, especially PNP bipolar of the Pplus/Nwell one, can help in sinking current during the ESD event.

3) Diodes capacitance must be taken into account in particular designs (RF systems).

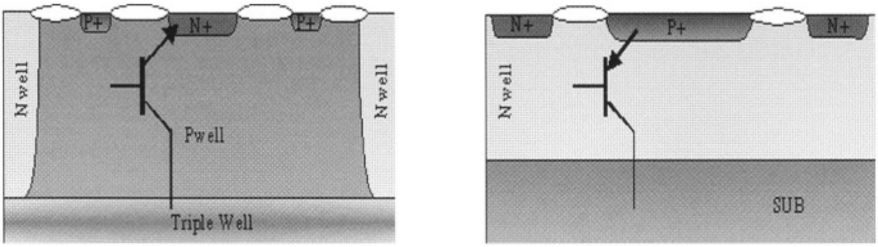


Figure 20: Cross section of Nplus/Pwell (diode between pad and ground line) and Pplus/Nwell (diode between pad and supply line) diodes with parasitic elements

ESD current in a common rail architecture is forced to pass through two forward biased diodes (one in case of ESD discharge vs. ground) and through the central protection clamp. A possible parallel path to the current could be activated if the pad voltage exceeds the diode breakdown value (diode working in breakdown region and a forward biased diode), but particular care has to be taken in the rail design in order to avoid this situation since diodes are very weak in breakdown region.

Considering, as an example, the case of a positive ESD discharge between a generic pad and the ground one: when the "in the pad" architecture is used, pad voltage is fixed by the protection characteristic plus the metal resistance to ground. On the other hand, in a common rail approach:

$$V_{PAD}=V_{BE}\,(I_{ESD})+R_{ON\_diode}*I_{ESD}+R_{METAL}*I_{ESD}/(\Box_{PNP}!\ V_{CLAMP}(I_{ESD}/(\Box_{PNP}!$$

.!

Diodes and central clamp has to be properly sized in order to limit the voltage below "dangerous" value fixed by the oxide thickness and by the diode breakdown. PNP parasitic bipolar transistor has to be properly characterized in the ESD current range since it plays a meaningful role sinking part of the current.

In figure 21 HBM robustness as a function of diode perimeter is reported for BCD6 process. Diode capability seems to scale pretty well with the diode perimeter: 30 λm seems to be the minimum value in order to guarantee 2kV.

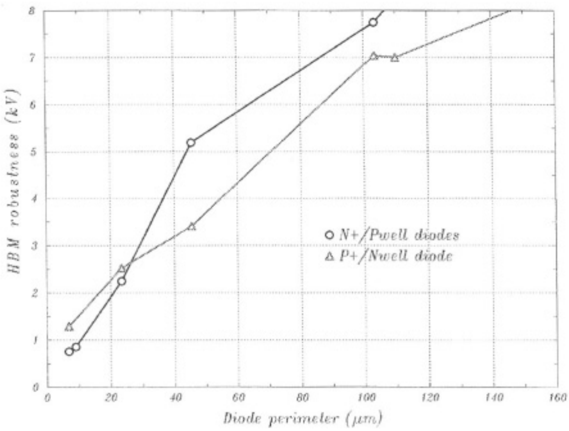


Fig 21: BCD6 diodes HBM robustness

Capacitance values may be reduced by one order of magnitude: in figure 22 diodes Nplus/Pwell Ron and capacitance are plotted as a function of its perimeter. Values lower than 100fF may be achieved still maintaining 2 kV HBM capability.

Chip size reduction may be also accomplished in case of large pin counts ICs.

Central clamp may be realized again using a self triggered lateral NPN properly sized: in fact even if a smaller current flows through this structure compared to the "in the pad" case (depending on the bipolar gain) the voltage drop across this structure has to be lower in order to compensate the ohmic drop across the routing metal line.

Dynamic clamp may be used with this architecture [12][13] being the clamp between pretty stable lines.

Common rail approach allows reaching high ESD ruggedness levels with a minimum impact on the die size due to the possibility to place only one large central clamp or to distribute it along the circuit periphery.

However when large robustness are required (large ESD current) proper sizing of diodes and central clamp cannot be done without taking into account the connection between the input pad and the central clamp. In fact large voltage drop across the metal routing line may lead to an increase of the pads voltage causing unacceptable parameters shift before inducing catastrophic damage.

In figure 23 threshold voltage shift measured on NMOS input buffer pull-down is reported as a function of metal series resistance present between the pad and a central clamp. This is the only damage caused by the ESD discharge: in fact both clamp and diodes do not reveal any leakage increase even after 8 kV HBM discharge.

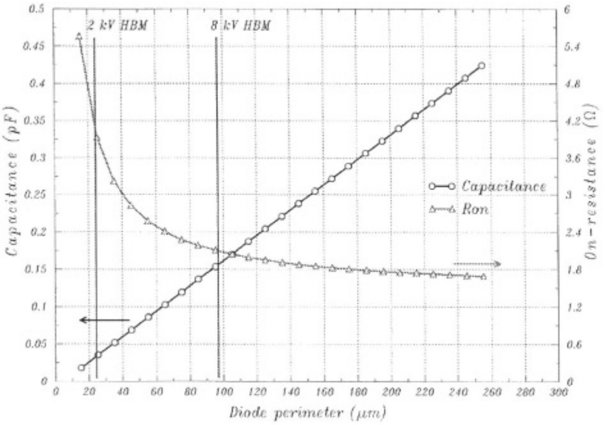


Figure 22: Nplus/Pwell diodes Ron vs. capacitance. Data are taken from BCD6 process

Threshold voltage shift is caused by the charge injection occurring during the ESD event induced by the pad voltage. Even if the detected shifts are pretty low and they would not cause any problem in many logic applications, the charge injected may have meaningful impact on oxide reliability performance.

From the figure it is evident that resistance value for good ESD performance must be below 1 ohm: thick metal level always present in smart power technologies does help in achieving this value without enlarging too much the metal strips but this could impose severe limit to the migration of the diodes' couple under the pad.

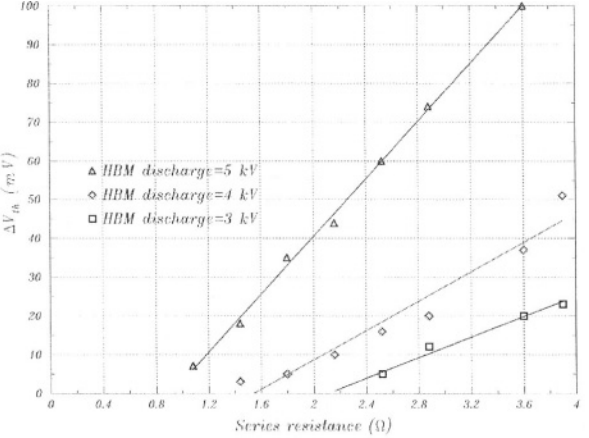


Figure 23: Voltage shifts in input buffer NMOS due to high metal resistance in a common rail architecture

Output stages

It is possible to protect output stages using the same architectures and protections previously analyzed for input pads. There are, however, some aspects that generally make difficult to apply exactly the same considerations above reported.

Series resistances in output stages may limit severely the current: only small resistance values are allowed. This make useless the use of the secondary element (see figure 8 a and b) for voltage clamping close to the real drain junction.

In "Open drain" configurations, where pull down NMOS alone is present, a diode connecting the NMOS drain to the supply is not allowed since a current path between two different supply lines (internal and external) may be created depending on the ICs' power up sequence (fail safe condition): the use of "in the pad" architecture protection network is therefore mandatory for these pads.

Moreover output stages have some peculiarities that distinguish them from input ones. First of all, the ESD event does not hit an oxide, but the external world comes in contact directly with a silicon junction. In the case of a common output buffer, the common drain is connected to the pad to be protected.

It can be observed that NMOS pull down device presents the same parasitic NPN bipolar already analyzed when GG/GCNMOS have been introduced. Basic difference is that normally gate terminal is driven by internal circuits and therefore should be considered as a "pseudo" floating terminal during the ESD event as it happens in the GCNMOS shown in figure 16. Hence this transistor may offer a parallel path to the current during the ESD event and in some cases (or for a short period of time) it may shunt the whole ESD current.

Hence particular care should be taken in drawing these transistors for a robust design since they would require particular tricks for improving their ESD robustness minimizing the current capability loss.

When large output transistor are used, protection may not be necessary since the device itself will be able to sink the whole current without damage.

When a protection is introduced (medium size transistors) it normally works in parallel with the transistors themselves. Depending on the turn on time of the protection structure and on its size the current flow through the output stages may become meaningful. It's therefore mandatory to adopt some of the solutions already shown for the protection robustness in the transistors' design. Increase drain side resistance by keeping the contacts "far" from the gate will improve current uniformity avoiding current localization. Contact spacing can improve noticeably HBM robustness (figure 24). An output buffer protected "in the pad" with a GCMOS structure, failing @ 4 kV when tested stand alone, would fail dramatically below 2 kV. Failure is localized in the pull down NMOS (Figure 25):

the poor drain resistance does not prevent current focalization that may cause silicon melted filament.

When however the output devices are correctly designed, overall ESD robustness may become larger than the one of the stand alone protection since the buffer may sink part of the ESD current.

At the same time, when a common rail approach is adopted the intrinsic body-drain diodes of the output buffer works in parallel with the ESD ones (when the two ESD lines coincide with supply and ground ones). Hence it is possible to downsize (or even eliminate) ESD diodes: transistors body connections have to be properly sized in order to carry ESD current peaks.

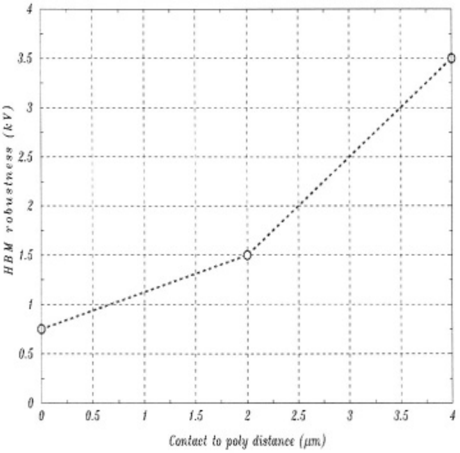


Figure 24: BCD6 3.3V output stage robustness dependence on poly-to-contact distance. Protection structure is a Gate coupled MOS. Protection stand alone failure voltage: 4 kV

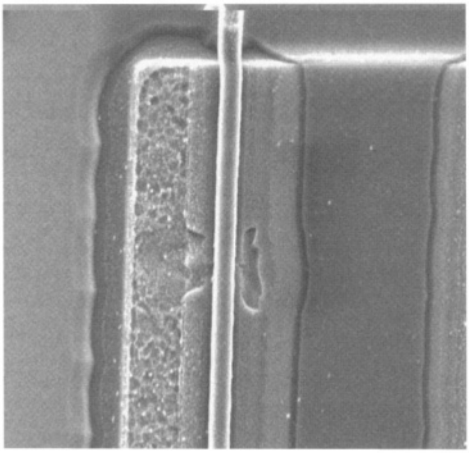


Figure 25: BCD6 3.3V output NMOS failed for a not robust design.

4.2. High voltage circuitry

The design of high voltage (from 16 to 80V) ESD protection devices in Smart Power technologies is a challenging activity since new and different constraints have to be considered.

As already reported in this article the concept of "ESD window" fixes very severe requirements, which must be fulfilled by the ESD protection structures: trigger voltage must increase accordingly to the operating conditions. Due to the high voltage level, large power dissipation occurs in the protection during the ESD event.

At the same time the safe window for ESD structure operation does not increase with the voltage class and since junctions' doping level has to decrease in order to improve voltage capability it's quite easy to imagine that High Voltage ESD structure electrical resistance will become larger so leading to worsen performances.

Static and dynamic solutions will be analyzed in the next sessions: both solutions are in fact widely used and are reported in literature.

As it will become clear an unique optimum solution may not be identified since the "high voltage" application field is very wide and is becoming wider in new generations where for example 5 V class can be already considered as part of it.

Dedicated solutions are often generated for peculiar applications: this makes very difficult to standardize a protection strategy.

Unique common starting point is the use of DMOS transistors, both lateral and vertical, for the realization of high voltage power switch realization: availability of this kind of transistors, that exhibits large current capability withstanding high voltage, may offer possible solutions for ESD protection design.

Recently advanced Smart power processes realized in a pure CMOS substrate without any dedicated step for high performance bipolar realization (heavily doped buried layer and sinkers, dedicated epi layer) have been used [4] for motor driver realization.

Cost and manufacturability push in that direction but ESD design and parasitic handling is becoming even more challenging!

4.2.1. In the pad solution: a static approach

High voltage self triggered bipolar devices arc probably the most widely used solution [20][26][28]. SCR with large blocking voltage and modified LDMOS architectures with an integrated SCR structure have been proposed [42].

One single bipolar NPN device properly designed for high voltage operation is commonly used. It's also possible, however, to use a chain of low voltage devices provided that trigger and holding voltages are compatible with the ESD process window.

Working mechanism of high voltage bipolar devices is the same of the low voltage ones: the bipolar is turned on by the avalanche hole current and a sudden voltage drop is recorded in the I-V characteristic.

Protection electrical resistance in snap-back region does depend on the collector region doping level: low doping level are mandatory in order to achieve large breakdown value, hence large resistance values have to be expected unless heavily doped buried layer and sinker are available.

TLP robustness level in the 20 mA/λm [28] range are reported for 0.35 λm smart power process with sinker and buried layer regions. In [28] the impact of the sinker presence is highlighted: protection current capability seems quite to be doubled when sinker is used. Such results do confirm what reported in [26]: vertical bipolar transistors are normally more ESD rugged than pure lateral bipolar devices.

As already reported in the previous session this is not always true at least for the low voltage class even if probably it becomes very difficult to minimize the power dissipation (electrical resistance) in a pure lateral high voltage bipolar structure.

Beside the use of high conductive layer, another key factor for the realization of robust structure is the holding voltage. Keeping this parameter as low as possible helps in the minimization of the power dissipation. It may happen that good ESD performances are obtained with low holding voltage (V_h φ 20 V) that is not compatible with the above defined ESD window. It's therefore very important to be able to tune independently trigger and holding voltage [33]. In this way protection design may be optimized according to the voltage class and to the performance of the device that has to be protected. It's in fact important to keep always in mind not only the solution itself but also what it has to be protected. For this reason close synergy between device and ESD structure development can be a key factor for the overall quality improvement.

In figure 26 a clear proof of what stated above is reported. Low voltage bipolar NPN chain is used as a protection element for an high voltage pad. A small size LDMOS drain is directly tied to the pad. In the figure TLP characteristic of the protection structure (NPN chain) and of the pad (NPN chain and monitor) are reported. Solid vertical lines fix the ESD window limited by the transistor breakdown (upper limit) and by the circuit maximum operating voltage (lower limit: it's reported the printer motor drivers' case).

It's evident that the protection works properly up to 1.3 A and its holding voltage guarantees safe circuit operation being larger than the lower ESD window limit (trigger voltage is not evident due to the scale resolution but its value is roughly 40V). Unfortunately, however, the large resistance after the snap back makes the pad voltage larger than the device breakdown, leading to the premature failing of the monitor LDMOS.

Such problems may be solved only through an enlargement of the protection structure or thanks to an enlargement of the ESD protection window that can be achieved thanks to a new LDMOS design structure (increasing device BV).

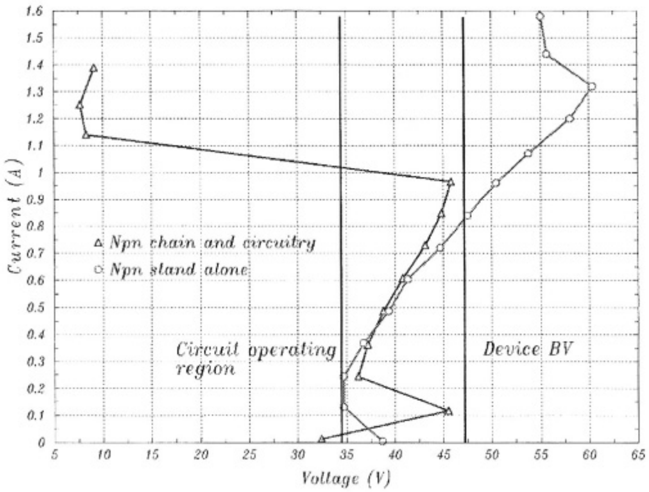


Figure 26: Ron of a chain of NPN bipolars is too high and causes monitor circuit failure

4.2.2. Common rail solution: a dynamic approach

As reported above it seems that vertical operation is very helpful in achieving good ESD performance when bipolar structures are used. So doing it's quite difficult to imagine a scaling of the protection size in advanced technologies. Current capability will be in fact fixed by the emitter area more than by its perimeter so that it would be very difficult to imagine any size reduction being the current fixed by the ESD requisite. Moreover lateral dimension arc actually fixed by voltage capability and it is not expected any change in future at least for present applications. Finally new processes, like some advanced smart power ones and SOI, are actually conceived without bipolar dedicated steps and it would be very interesting to investigate different solutions for ESD protection design.
Moving to a common rail architecture (see figure5), large area dynamic (or active) clamps [34] may offer a possible solution.
The unique large clamp is put between high voltage line and low voltage line (ground) and several pads discharge their current to this unique clamp (two diodes per pad are necessary).

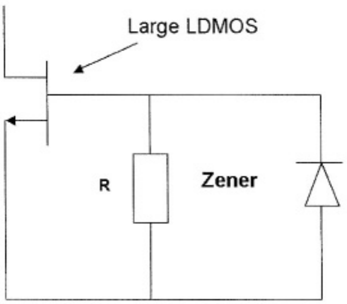


Figure 27: Standard dynamic clamp simplified schematic.

A simple proposed dynamic clamp is reported in figure 27. It basically consists of a “large” area LDMOS device whose gate is kept “floating” during the ESD event by a resistance placed between the gate and the source.

When an ESD pulse is applied, this transistor is switched on capacitively: thanks to its particular design in fact, LDMOS gate to drain capacitance may be quite larger than standard NMOS transistors, thus increasing the coupling. At the end of the pulse the transistor is switched off with the time constant imposed by the resistor value that has to be properly designed. Clearly the gate discharge time constant has to be longer than the ESD pulse but too long time constant may lead to clamp failure since LDMOS maximum energy capability would be exceeded.

In figure 28 dynamic I-V characteristic obtained through TLP testing is reported for a 20 V LDMOS processed in BCD6 technology. Clamp width is 11 mm.

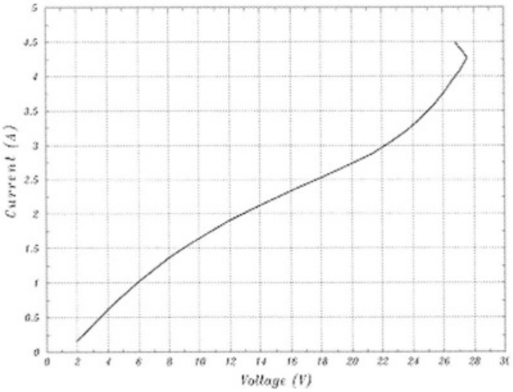


Figure 28: BCD6 20V standard clamp I-V characteristic

Large coupling between drain and gate are very attractive since they allow large current discharge without large voltage generation across the clamp.

The correct dimensioning of the central clamp will be done according to the total current which must be shunted to ground and to the total voltage drop across the clamp, which should be kept as low as possible in order to reduce the stress on the protected stages.

LDMOS current capability is fixed by the channel length and it's going to increase with new technologies together with the channel density (channel width integrated by unit area). Even if actual current capability levels are much lower than the values achieved with bipolar devices, values in the range of 0.4-0.5 mA/λm are achieved in a 0.35 λm technology with 20 V LDMOS devices, due to the very large channel density. Therefore area consumption is expected to become comparable. When in fact the current capability is normalized to the ESD protection area consumption the reduced LDMOS cell pitch allows to reach quite large normalized values: BCD6 20 V LDMOS device specific current capability is roughly 120 A/mm^2.

In figure 29 BCD6 20V active clamps HBM failure level as a function of clamp width is reported. Good scaling with the electrical width is observed: curve slope is roughly 410 V/mm that is pretty lower compared to the theoretical expected value evaluated by the device current capability (600 V/mm).

Probably the device is not turned on uniformly as suggested in [35]: the presence of multiple gates in parallel makes very difficult to ensure the uniform biasing of whole channel perimeter.

Experimental evidence of gate potential non-uniformity is provided in [36] where even a gate grounded LDMOS device seems to turn on during locally during the ESD event.

It may also be possible, however, that the reduced slope is caused by a poor gate to drain coupling: drain voltage may exceed breakdown value while gate to source voltage is still below maximum value.

However good HBM robustness scaling with the clamp total width has been obtained: this analysis can help in the right dimensioning of the clamp depending on the desired robustness level.

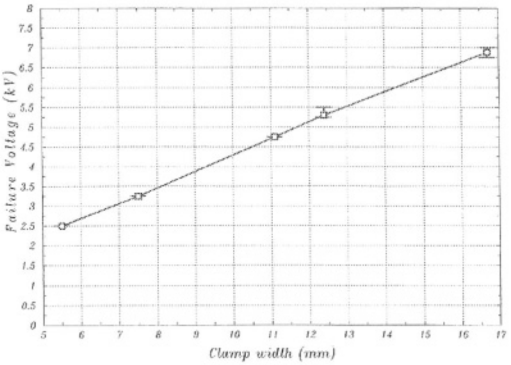


Figure 29: 20 V LDMOS active clamp HBM failure voltage as a function of the clamp width.

Moreover the above figure may give another very useful information: the minimum power LDMOS size that does not require any protection. In practical applications large LDMOS power switches are mandatory for minimizing power loss: they can be considered as self protected if their dimension is larger than the value reported in the above figure. This is true if the power gate driver circuit does not discharge the gate during the ESD event: in such case proper design modifications are mandatory for reaching “safe” ESD robustness level [37].

Variations of the simple structure presented above have been presented in literature. In order to increase drain to gate coupling a smaller LDMOS (same voltage capability) has been put between the two terminals in the so called “Darlington” configuration [38]. TLP characterization did confirm that same current is shunted through the clamp with a significative minor voltage drop across it. HBM robustness shows significative improvements thanks probably to the reduced power dissipation during the ESD event.

By inserting a zener chain between gate and drain no significative improvements either in TLP characteristics nor in terms of HBM robustness have been obtained [38]: from TLP analysis it seems quite evident that the clamp is activated dynamically and the zener chain does not affect the clamping voltage. However this solution may be very useful in real ICs design where the high voltage rail normally is the supply line. Large capacitance associated with the supply line may slow the pulse rise time and the clamp may not be turned on: the presence of the zener chain ensures that the large LDMOS is switched on even statically when the critical voltage is reached [35].

Once good ESD performances are reached with appropriate clamp sizing, problems may arise in the protected circuitry. This problem is not often treated in literature especially in the case of high voltage pads. The active clamp solution may become critical when a "small" size high voltage LDMOS device has to be protected (example: an high voltage output device with reduced current capability). In this case, in fact, the circuit that has to be protected is identical to the protection itself: one LDMOS transistor with a potentially floating gate. This transistor is usually smaller and directly tied to the pad whereas the protection is larger and connected to the pad through a diode. Normally it can turn on immediately shunting the current at least during the first part of the discharge.

In figure 30 this situation is reported for a 20 V class pad: small size LDMOS (Width= 500 λm) is directly tied to the pad that is protected using a common rail approach with a standard active clamp between the lines. In the figure the pad HBM robustness level is reported as a function of the large clamp width.

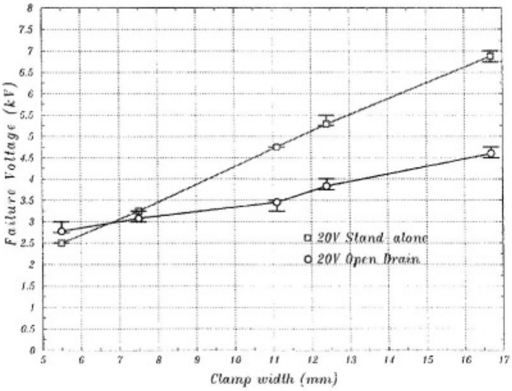


Figure 30: Open drain pad HBM robustness level as a function of the central clamp width. Stand alone robustness is reported too.

In the same picture the results already shown in figure 29 and obtained on a stand alone clamp are reported too. As evident, when a monitor device is introduced trying to simulate simple ICs case, the situation changes dramatically. Even if no fail has been detected in the protection structure, the monitor device fails at much lower level.

Significative improvements may be achieved by making the clamp faster than the protected device: Darlington solution, for example, seems to be very effective in achieving this task (Figure 31). Increased robustness slope (520 V/mm vs. 400 V/mm) may be caused by an improved turn on uniformity due to the insertion of an external capacitive load between gate and drain or by the increased drain to gate coupling.

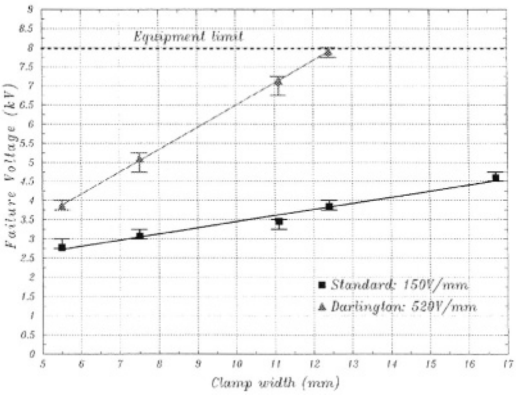


Figure 31: Open drain pad HBM robustness level as a function of the central clamp width. Darlington and standard clamp structures are compared. 8 kV is the equipment limit.

Current uniformity in the large LDMOS devices and premature turn on of the protected transistors are the key issues that has to be faced when dynamic ESD protection strategy is adopted.

Moving to higher voltage classes (larger than 45 V) the situation may become even more critical. The use of lightly doped drain extension, mandatory to satisfy the voltage requirements, makes LDMOS device very weak in the high current regime. Large multiplication current, generated when the current through the transistor is large (i.e. biased with large V_{gs}) may induce premature snap back [39].

Due to the particular body contact topology peculiar of LDMOS device, preferential localized snap back regions may be identified [38][40]. Current focalization after snap back will lead to the device's failure.

Even if salicide technique helps in avoiding body contacts non-uniformity, the body current dependence on the gate to source local bias may still induce localized snap back.

In [41] large HBM failure voltage spread is reported when 40 V Darlington clamp is analyzed: this behavior is normally attributed to localized failures that have been evidenced by an extensive failure analysis activity. Moreover no dependence on large clamp LDMOS width was detected when open drain pads has been tested. Due to their intrinsic weakness the small size devices immediately fails during the first part of the ESD event when they probably turn on.

Significative improvements in the pad robustness may be achieved decreasing the clamp turn on time working again on the drain gate coupling.

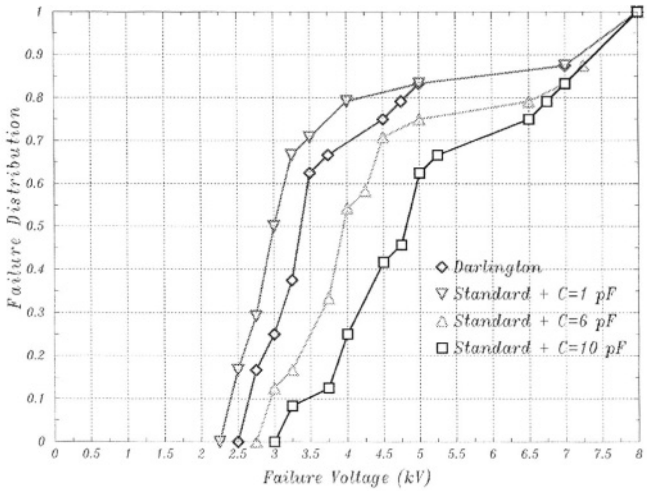


Figure 32: 40 V open drain pad HBM robustness level cumulative distribution. HV linear capacitor (MOM) is introduced between gate drain (data for BCD6 process).

In figure 32 the effect of the introduction of an high voltage capacitor between drain and gate is shown. It seems that a significative reduction of the low failure voltage population is achieved increasing the capacitance value.

Similar effect may be reached working on the robustness of the small size device. In figure 33 a comparison between two pads sharing the same protection structure, a Darlington clamp, realized with different LDMOS devices, 40 and 70 V LDMOS, is reported. The reduction in the low failure level percentage can be attributed only to the increased small size device robustness.

The proposed solutions (figure 32 and 33) are both effective in improving the ESD robustness but the data spread is still pretty high compared to what it's possible to achieve in lower voltage class (see data for 20 V LDMOS in [41] for example).

It's probably still possible to reduce the data spread by working on the large area clamp device's uniformity: a novel architecture, still under investigation, has shown excellent preliminary results.

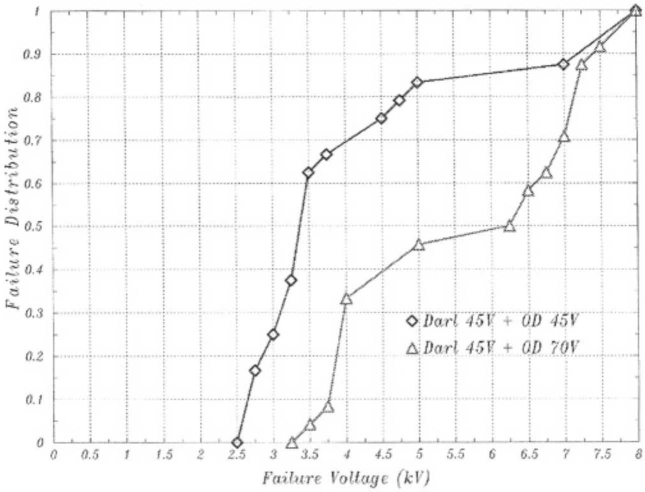


Figure 33: 40 V open drain pad HBM robustness level cumulative distribution. Different protected devices are compared: 40 V vs 70 V small size LDMOS.(data for BCD6 process)

5. Conclusions

Electronic devices penetration in many new application fields does often require a significative increase of the modern ICs quality: ESD is still one of the major concern that electronic industry has to face for achieving this task.

Costs associated with new mask set realization and faster time to market often impose first design to be successfull. Circuit simulation may be very helpful for the identification of possible design marginalities. Modeling of electronic devices behaviour in the ESD pulse time regime is mandatory: new TLP equipments make possible the extraction but due to the strong non linearities of the protection structures and due to the complexity of the simulations (true expecially for CDM), they still require long CPU time.

Thanks to the efforts spent in the recent years and to the new analysis tools availability many steps forward have been made for the understanding of ESD related phenomena.

However modern technologies and increased circuit complexity continuosly introduce new issues that ESD protection designers have to face. Gate oxide scaling seems to be the most critical topic that will have to be faced in advanced CMOS platforms. Latent damage caused by the electrical stress during the ESD event may cause severe reliability problems difficult to be screened at circuit testing level.

Shorter time to market is required by new market conditions, hence ESD protection design has to be carefully considered since the beginning of a new process development.TCAD simulation tools are currently widely used for a qualitative understanding of involved physic mechanisms: solid and reliable 3 D tools are probably mandatory for a real predictive analysis of ESD protection behavior since many failures are still caused by localized phenomena.

Within this scenario, smart power processes complexity is growing accordingly to new products requirements, but high voltage ratings arc not scaling accordingly.

Large power dissipation during the ESD event in high voltage pads has to be handled within an advanced CMOS like platform: thinner junctions, thinner gate oxides and pure lateral devices.

LDMOS based dynamic clamps seem to be a robust solution for the high voltage pad protection realization since they do not require dedicated process steps and they have already been used with success in many BCD process platforms.

References

[1] Amerasekera A.,Duvvury C., “ESD in silicon integrated circuits”, Wiley, 1995

[2] Vinson J.E., Liou J. J., “Electrostatic Discharge in Semiconductor Devices: An Overview”, IEEE proc., vol. 86, n.2, 1998

[3] Wagner R. G. et al., “Extent and Cost of EOS/ESD Damage in an IC Manufacturing Process”, EOS/ESD proc. Symp., 1993

[4] Moscatelli A. et al., “LDMOS Implementation in a 0.35μm BCD technology (BCD6)”, ISPSD proc., 2000

[5] Parthasarathy V. et al., “A 0.25μm CMOS based 70V smart power technology with deep trench for high voltage isolation”, IEDM proc., 2002

[6] Contiero C. et al, “Roadmap Differentiation and Emerging Trends in BCD Technology”, ESSDERC proc, 2002

[7] ESD STM5.1-1998, ESD association standard test method

[8] Verhaege K. et al, “Influence of Tester, Test Method, and Device Type on CDM ESD Testing”, IEEE trans. on components, packaging. And manufacturing technology, vol.18, n. 2, 1995

[9] Maloney T. and Khurana N., “Transmission line pulsing techniques for circuit modeling of ESD Phenomena”, EOS/ESD symp. proc., 1985

[10] Gieser H. et al, “Very-fast transmission line pulse of integrated structures and the charged device model”, EOS/ESD symp. proc, 1996

[11] Richier C. et al., "Investigation on Different ESD Protection Strategies Devoted to 3.3V RF Applications (2 Ghz) in a 0.18 μm CMOS Process", EOS/ESD symp. proc., 2000

[12] Worley E. R. et al., "Sub-micron chip ESD protection schemes which avoid avalanching junctions", EOS/ESD symp. proc., 1995

[13] Torres A. C. et al., "Modular, Portable, and Easily Simulated ESD Protection Networks for Advanced CMOS Technologies", EOS/ESD symp. proc., 2001

[14] Duvvury C., "ESD protection devices issues for IC designs", IEEE Custom Integrated Circuits Conference proc., 2001

[15] Duvvury C., "A Fail-safe ESD protection circuit with 230 fF Linear Capacitance for High Speed/High Precision 0.18 μm CMOS I/O application", EOS/ESD symp proc, 2002

[16] Groeseneken G., Hot Carrier Degradation and ESD in Submicron CMOS Technologies: How do they Interact?“, EOS/ESD symp. proc., 2000

[17] Wu J. et al., "Breakdown and Latent Damage of Ultra-Thin Gate Oxides under ESD Stress Conditions", EOS/ESD symp. proc., 2000

[18] Stadler W. et al., "Device Simulation and backside Laser Interferometry – Powerful Tools for ESD Protection Development", ESREF proc., 2002

[19] Kassakian F.G. et al., "The future of Electronics in Automobiles", ISPSD proc., 2001

[20] Moens P. et al., "Future Trends in Intelligent Interface Technologies for 42V Battery Automotive Applications", ESSDERC proc., 2002

[21] Gauthier R. et al., "Evaluation of Diode-Based and NMOS/*Lnpn*-Based ESD Protection Strategies in a Triple Gate Oxide Thickness 0.13 μmCMOS Logic Technology", EOS/ESD symp. proc., 2001

[22] Kunz K. et al., "5V Tolerant Fail-Safe ESD Solutions for 0.18 μm Logic CMOS Process", EOS/ESD symp . proc., 2001

[23] Ker M.D. et al., "How to Safely Apply the LVTSCR for CMOS Whole-Chip ESD Protection without being Accidentally Triggered on", EOS/ESD symp. proc., 1998

[24] Meneghesso G. et al. "ESD robustness of smart-power protection structures evaluated by means of HBM and TLP tests", IRPS proc., 2000

[25] Amerasekera A. et al., "The impact of Technology Scaling on ESD Robustness and Protection Circuit Design", IEEE trans. on components, packaging. And manufacturing technology, vol.18, n. 2, 1995

[26] De Heyn V. et al., "Design and Analysis of New Protection Structures for Smart Power Technology with Controlled Trigger and Holding Voltage", IRPS proc., 2001

[27] Blaho M. et al., "Experimental and simulation analysis of a BCD ESD protection element under the DC and TLP stress conditions", ESREF proc., 2002
[28] De Heyn V. et al., "Effect of the n^+ Sinker in Self-Triggering Bipolar ESD Protection Structures", EOS/ESD symp. proc., 2002
[29] Depetro R. et al., "Experimental Analysis and Electro-Thermal Simulation of Low- and High- voltage ESD Protection Bipolar Devices in a Silicon-On-Insulator Bipolar-CMOS-DMOS Technology", EOS/ESD symp. proc., 2001
[30] Verhaege K. G. et al., "Wafer Cost Reduction Through Design of High-Performance Fully Silicided ESD Devices", EOS/ESD symp. proc., 2000
[31] Pogany D. et al., "Study of Trigger Instabilities in Smart Power Technology ESD Protection Devices Using a Laser Interferometric Thermal Mapping Technique", EOS/ESD symp. proc., 2001
[32] Salling C. et al., "Development of Substrate-Pumped NMOS Protections for a 0.13 μm technology", EOS/ESD symp. proc., 2001
[33] Pendharkar S. et al.,"ESD robust bipolar transistors with variable trigger and sustaining voltages", ISPSD proc., 2003
[34] Meneghesso G. et al., "Overstress and Electrostatic Discharge in CMOS and BCD Integrated Circuits", ESREF proc., 2000
[35] Duvvury C. et al., "Lateral DMOS Design for ESD Robustness", IEDM proc. 1997
[36] Besse P. et al., "Investigation for a Smart Power and Self-Protected Device Under ESD Stress through Geometry and Design Considerations For Automotive Applications", EOS/ESD symp.proc., 2002
[37] Van Zwol J. et al., "ESD Protection by Keep-On Design for a 550V Fluorescent Lamp Control IC With Integrated LDMOS Power Stage", EOS/ESD symp. proc., 2002
[38] Sponton L. et al., " ESD protection structures for BCD5 smart power technologies", ESREF proc. 2001
[39] Ludikhuize A.W., "Kirk Effect limitations in High Voltage ICs", ISPSD proc., 1994
[40] Mergens M. et al., "Analysis and Compact Modeling of Lateral DMOS Power Devices Under ESD Stress Conditions", EOS/ESD symp. proc., 1999
[41] Sponton L. et al., "ESD protection structures for 20V and 40V power supply suitable for BCD6 smart power technology", ESREF proc., 2002
[42] Pendharkar S. et al., "SCR-LDMOS – A Novel LDMOS Device With ESD Robustness", ISPSD proc., 2000

RF-ESD Co-Design for High Performance CMOS LNAs

Paul Leroux, Michiel Steyaert
K.U.Leuven, ESAT-MICAS
Kasteelpark Arenberg 10
B-3001 Heverlee, Belgium

Abstract

This paper fits within the framework of recent research on the use of pure CMOS, rather than bipolar or BiCMOS technologies for RF front-ends. This paper focuses on the Low Noise Amplifier which is commonly the first building block in a wireless receiver. Since the LNA input pin connects to the outside world, it is sensitive for Electrostatic Discharges (ESD). Although this is a critical issue, very few LNA papers [17][18][23] have been published with ESD-protection results This paper gives guidelines for a rigorous RF-ESD co-design for high performance CMOS LNAs. In this work, two important LNA topologies will be discussed and different strategies for RF-ESD co-design will be explained.

1. Introduction

One of the main bottlenecks for introducing CMOS RF circuits to the market is their susceptibility to ESD. It is due to both gate oxide breakdown and junction degradation related problems, caused by the decreased oxide thickness and increased doping levels in the scaled down technologies [13]. The ESD problems are still increased by the tight design window for the high performance RF circuits, not allowing large ESD devices to be used as protection elements [20]. Most CMOS ESD-protection structures (e.g. as they are used in digital CMOS) have parasitics that are detrimental for the LNA performance. They commonly feature two large clamping devices with a current limiting resistor in between. The resistance added at the input (up to a few hundred ohms), would be detrimental for the noise figure of the LNA. The introduced parasitic input capacitance also has a serious influence on the LNA performance. This influence will be explained in section 2.

Different ESD protection elements can be found in open literature, diodes, silicon controlled rectifiers (SCR), grounded gate NMOS (GGNMOS), gate coupled NMOS (GCNMOS), ... Also different ESD-protection tests and standards are used in industry. Among them, the human body model (HBM), machine model (MM), the charged device model (CDM). Another commonly used test is the transmission line pulse (TLP). One of the most commonly used standards HBM test. It consists of a charged capacitance of 100pF, that is

A. van Roermund et al. (eds.), Analog Circuit Design, 207-226.
© 2003 *Kluwer Academic Publishers. Printed in the Netherlands.*

discharged through a 1.5kΩ resistor, connected to the DUT. The amount of protection is indicated by the HBM voltage the circuit is able to withstand. The standard level of protection for an IC is 2kV. The maximum discharge current through the DUT is in the order of 0.7A/kV. Hence, a protection level of 2kV means the circuit should be able to carry 1.4A while limiting the voltage at the input to a safe level. This value will be used to discuss the design methodology. However note that a different standard does not change the discussion.

A lot of different types of LNAs are known and have been published in open literature. A distinction is made between wideband or baseband amplifiers [5][21] and narrowband or passband amplifiers. In what follows only the latter will be discussed since we are dealing with wireless systems in which the signal power is concentrated around the carrier frequency. The most interesting and widely used topology for narrowband LNAs is based on a single transistor common source (CS) amplifier. It is most interesting since it allows a very low NF and high gain which is mandatory for many of today's wireless receivers requiring a good sensitivity (GPS, GSM, DCS 1800, UMTS, ...). An RF-ESD co-design strategy for this topology will be discussed. Results of this design strategy are demonstrated with the measurements of an LNA for a portable GPS receiver [23]. Section 6 discusses another ESD-protection strategy more attractive for higher frequency LNAs, operating at more than 5GHz. It is based on the use of an on-chip inductor as a signal splitter [22]. Both calculations and simulation results are shown. Yet another ESD-protection strategy based on a Π-matching network is discussed in section 7. In section 8, we will take a closer look at the common gate LNA and show when this topology is more interesting.

2. Basic operation of the common source LNA

Figure 1 depicts the input of an inductively degenerated, common source LNA. Consider first the ideal case where no parasitic input capacitance, C_p, is present, V_S and R_S represent the signal source that feeds the LNA. This is the output of either the antenna or a bandpass channel select filter. In both cases the impedance of the source is 50Ω. As a consequence the LNA input should be matched to 50Ω in order to absorb as much signal power as possible and to avoid distortions in the filter characteristic due to incorrect termination. Since the common source amplifier ideally has no input resistance, only input capacitance, the source degeneration inductor L_S provides an 'artificial' resistive input impedance equal to $\omega_T.L_S$. The gate-source capacitance is tuned out by a gate inductor, L_g, in order to create a purely resistive input.

However, in real life, some parasitic input capacitance is always present. Part of it is coming from the input bonding pad, another part belongs to the gate-drain capacitance. A normal ESD-protection network adds even more capacitance to that. This capacitance, C_p, increases the noise figure of the LNA and decreases its gain. It also places an upper bound on the attainable input impedance values. If ωC_p is too high, it even becomes impossible to obtain an input match without an external matching circuit.

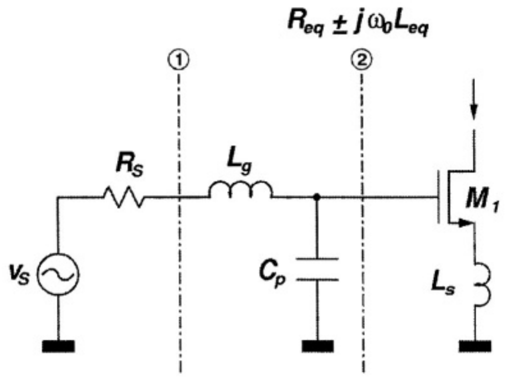


Figure 1 Influence of C_p on LNA performance.

It is instructive to define two reference planes 1 and 2 and to look at the different impedance levels at these nodes. Looking towards the source at reference plane ② in Figure 1, the gate sees an equivalent input resistance R_{eq} larger than 50Ω, and a series inductance L_{eq} slightly larger than L_g. They are given by:

$$R_{eq} = \frac{R_S}{\omega_0^2 C_p^2 R_S^2 + \left(1 - \omega_0^2 C_p L_g\right)^2} \tag{1}$$

$$L_{eq} = \frac{L_g - C_p\left(\omega_0^2 L_g^2 + R_S^2\right)}{\omega_0^2 C_p^2 R_S^2 + \left(1 - \omega_0^2 C_p L_g\right)^2} \tag{2}$$

The noise factor of the LNA may be shown to be approximately given by

$$F = 1 + \left(\frac{\omega_0}{\omega_T}\right)^2 \frac{\gamma}{\alpha} g_m R_{eq} + \frac{\alpha\delta.\left(1 - \left|c^2\right|\right)}{\kappa.g_m R_{eq}} + \left(\frac{\omega_0}{\omega_T}\right)^2 \frac{\gamma}{\alpha}\frac{2}{\kappa} \tag{3}$$

Notice that the noise factor features a term directly proportional to $g_m.R_{eq}$, coming from the classical drain noise, and a term inversely proportional to $g_m.R_{eq}$, coming from the non-quasistatic noise. Due to the presence of C_p and hence, the higher value for R_{eq}, the noise figure is primarily determined by the classical drain noise contribution. As the ESD-protection circuit even increases C_p, R_{eq} becomes larger and the noise figure increases further.

C_p also has an influence on power gain. In order to have a high power gain, a high squared output current ($|i_{out}|^2$) is required for a given available source power, P_{av}. Its value may be calculated as a function of R_{eq}:

$$|i_{out}|^2 = \left(\frac{\omega_T}{\omega_0}\right)^2 \frac{P_{av}}{R_{eq}} \tag{4}$$

where P_{av} is the available input power from the antenna. R_{in} is the real part of the impedance, seen to the right of reference plane 2. It is equal to R_{eq} for a perfect input match. The power gain G_t is then given by:

$$G_t = \frac{R_{load}}{4 \cdot R_{eq}} \left(\frac{\omega_T}{\omega_0}\right)^2 \tag{5}$$

If the value of C_p increases, R_{eq} becomes larger and the amount of output current decreases. Hence, a higher C_p implies a lower power gain.

The dependence of the LNA linearity, characterized by its IIP3, on the V_{GS}-V_T (V_{gt}), technology and input matching is shown in (6).

$$\begin{aligned} IIP3[\mathrm{dBm}] \approx & -16 + 10\,log\left(\frac{4}{3}\frac{V_{gt}(1+\Theta V_{gt})^2(2+\Theta V_{gt})}{\Theta}\right) \\ & + 20\,log\left(\omega_0 C_{gs}(R_{eq}+R_{in})\sqrt{\frac{50\Omega}{R_{eq}}}\right) \end{aligned} \tag{6}$$

The first part of this equation is derived from the well known MOS current equation under mobility degradation which is modeled by the factor Θ. For a transistor without matching section (only the first term in (6)), the IIP3 improves with increasing V_{GS}-V_T and deteriorates with deeper submicron technologies. For a matched transistor, the dependencies get more complex because the IIP3 decreases with decreasing equivalent source resistance (R_{eq}). The –16 term converts the units from $\mathrm{dBV_{amp}}$ to $\mathrm{dBm_{av}}$.

3. RF Design considerations

In this section the most important design considerations are discussed using a 1.57GHz LNA for a portable GPS receiver as a demonstrator. A simplified schematic of the LNA is shown in Figure 2. The biasing details are not shown. In this design, the input bonding wire is used as gate tuning inductor since the Q of an on-chip inductor would be to low resulting in a high noise figure. The input bonding pad is shielded from the substrate by grounding the bottom metal layer [17]. This drastically increases the Q of the bond pad capacitance and prevents noise coupling from the substrate. Furthermore, the pad capacitance is minimized by using only the top metal layer and octagonal layout. Limiting this capacitance is of extreme importance to create a sufficiently large headroom for the capacitance of the ESD devices. The details of the ESD-protection network will be discussed in the next section.

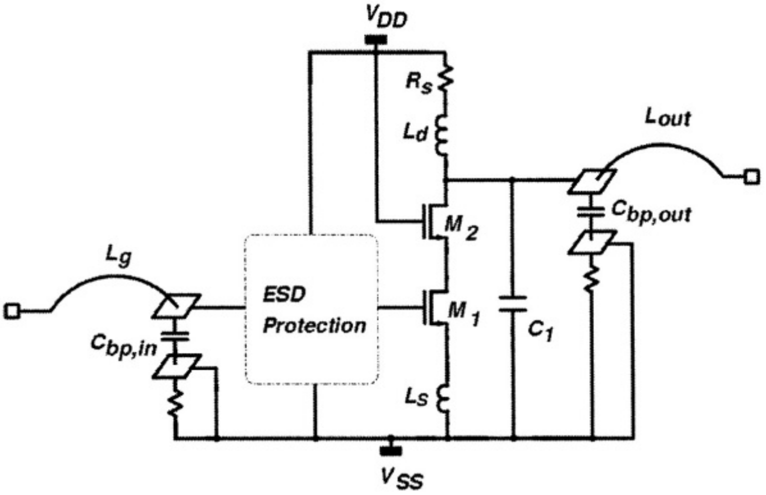


Figure 2: Simplified LNA schematic.

The source degeneration inductor, L_s, is implemented by parallel bonding wires. The load inductor, L_d, is realized on chip. Its series resistance is depicted as R_s and its parasitic capacitance to the patterned ground shield beneath it, is included in C_1. C_1 (together with the pad capacitance $C_{bp,out}$) and bonding wire L_{out} constitute the output matching network for obtaining a 50Ω output impedance.

Equations (1) to (6) have been used to describe the design space of the LNA and to visualize the influence of the equivalent source resistance (R_{eq}) and the transistor sizing on the noise figure, gain and linearity of the LNA. A higher value for R_{eq} is beneficial for the linearity but is bad for noise and gain performance. Since, for the GPS application, the latter constraints are more binding, R_{eq} is chosen sufficiently low. As R_{eq} increases with the amount of parasitic input capacitance (1) the capacitance available for the ESD-protection is limited. Moreover, the tolerated capacitance decreases with the operating frequency of the circuit.

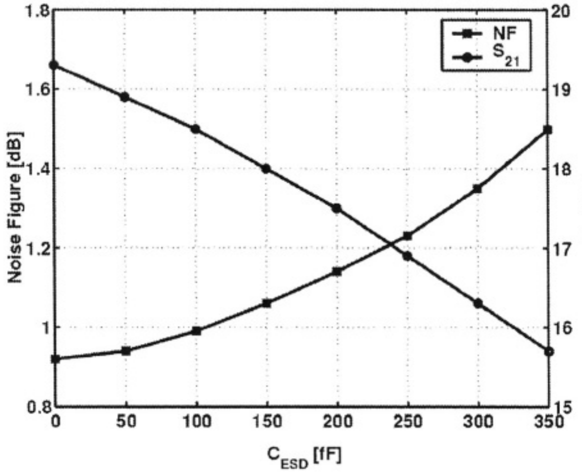


Figure 3: LNA Noise Figure and Power Gain as a function of ESD-capacitance.

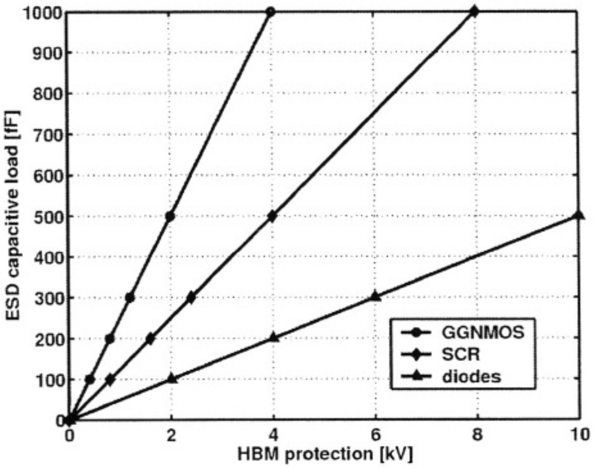


Figure 4: ESD-capacitance as a function of HBM protection voltage for different devices.

The LNA has been simulated and optimized both with and without the ESD-protection diodes. The MOS model that was used is level 49, BSIM3v3. Both the MOS transistors and the diodes were placed within a subcircuit to include extra parasitics that are not implemented in the BSIM3v3 model. This allowed us to model the Non-QuasiStatic effect of the channel charge buildup which has a significant influence on the RF-performance. Bulk and well resistances for both the diodes and transistors have been included since they influence both high-frequency and high-current behavior.

The noise figure and power gain are plotted vs. the capacitance of the ESD-protection diodes in Figure 3. The other contributors to the capacitance are also taken into account in these simulations but they are not included in the value of C_{ESD}. It is seen in this figure that the NF increases from 0.9dB to 1.5dB for a capacitance of 350fF. The power gain decreases from 19.3dB to 15.7dB. Beyond this capacitance value the circuit can no longer be matched to 50Ω without an external matching network. For higher frequencies this effect becomes more pronounced, i.e. the curves will be steeper and the cut-off capacitance becomes lower. These curves can be redrawn for different power budgets.

Table 1: Comparison between the simulated RF-performance of the LNA with and without input ESD-protection.

	Without ESD-protection	*With ESD-protection*
Power	9mW	9mW
Current	6mA	6mA
Supply voltage	1.5V	1.5V
S_{11} @ 1.57GHz	-14dB	-14dB
S_{21} @ 1.57GHz	19.3dB	18.4dB
S_{12}	< -30dB	<-30dB
S_{22} @ 1.57GHz	-23dB	-25dB
NF @ 1.57GHz	0.9dB	1.0dB
IIP3	-6.8dBm	-5.1dBm

Figure 4 plots the ESD-capacitance as a function of HBM protection voltage for different devices. Together with Figure 3 this plot allows the designer to choose the amount of ESD-protection required and tolerable for a specific application. In our design the ESD diodes have a capacitance of about 100fF yielding an ESD protection of 3kV HBM.

Table 1 shows a comparison between the main simulation results with and without ESD-protection. This table confirms the theoretical discussion given above. Adding the ESD-protection at the input reduces the power gain from 19.3dB to 18.4dB. The noise figure increases from 0.9dB to 1dB. The IIP3 improves from -6.8dBm to -5.1dBm.

4. ESD protection design considerations

The ESD protection network is shown in Figure 5. It consists of two diodes, D_1 and D_2, between the RF input and the power supply busses and of a stack of five diodes D_3-D_7 between V_{dd} and V_{ss}. The goal is to provide an explicit ESD discharge current path for all possible stress combinations between the input, V_{dd} and V_{ss}. The use of diodes was based on the fact that they are very efficient and robust ESD devices. Furthermore, their characteristics are fairly simple to model and simulate, allowing a reliable sizing of these devices. Since the LNA was designed for a fully integrated GPS receiver, where the output node of the LNA directly connects to the mixer input, no output ESD-protection is required.

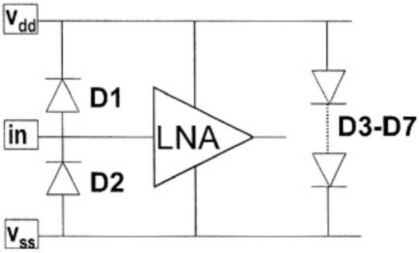


Figure 5: The P+Nwell diode based ESD protection network designed around the core RF-LNA circuit.

Key considerations in the D_1-D_2 input diodes design is their capacitance, and their high frequency and high current resistance. The first two affect the RF performance of the LNA (as shown earlier), the second influences the ESD performance of the circuit and, in particular, the bias developed on the input node/gate of the LNA during the ESD pulse. Both the capacitance and the resistance should be minimized.

The diodes can be laid out as one finger or multiple finger devices. It is a well known fact that the junction capacitance of the diode is related mostly to the bottom plate of the diode while the current flows mostly through the sidewall. Hence to minimize the capacitance and the resistance of the diode we need to minimize the area and maximize the perimeter. This clearly pleads in favor of a multiple finger structure. Moreover, a one finger diode may give current crowding at the corners, yielding dangerous hot spots where breakdown may initiate. These fingers may be realized either as squares or stripes. Stripes give the advantage of a very low resistance since the average distance from a p+ to n+ contact is minimal while still providing sufficient contacts per finger. On the other hand, current crowding and the resulting hot spots may occur at the small ends of the stripes. This problem may be somewhat alleviated by removing the contact closest to the end of the stripe in order to increase the resistance in that path preventing current crowding. When using squares, the average resistance will be a little higher than with stripes but a uniform current distribution is more easily obtained.

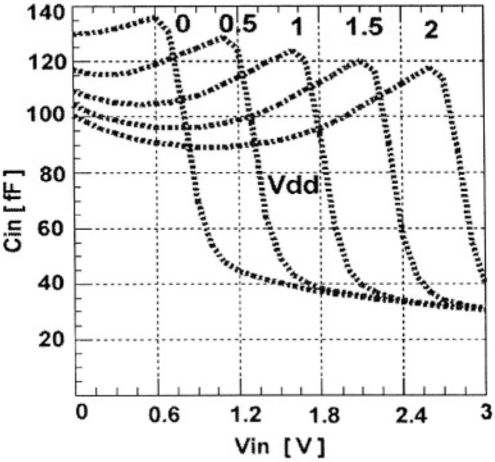


Figure 6: The simulated input capacitance contribution of D1 and D2 as a function of the applied input DC bias.

The chip discussed in section 5 uses a multiple finger structure of square diodes. The area of the squares was chosen as small as possible while keeping more than one via for each square. The area of one diode square in this design is $9\mu m^2$ giving a total area of $36\mu m^2$ for a total capacitance of 50fF per diode (one to V_{dd} and one to ground). Simulation of the low frequency input capacitance C_{in}, due to D_1 and D_2, as function of the applied DC input and V_{dd} biases is shown in Figure 6. Since the value of C_{in} is relatively constant around the DC operating point, it does not significantly degrade the linearity of the circuit.

The design of the diode string D_3-D_7 between the supply rails was based on the V_{dd} to V_{ss} leakage current specifications and the small high current resistance requirements. In an nwell CMOS technology a p+ nwell diode is actually a pnp transistor [1]. Hence, when you connect a series of diodes and some current is injected in the top diode, the current through the next diode is reduced by a factor $1+\beta$, where β is the current amplification of the parasitic bipolar transistor. As a consequence the chain of diodes will start to conduct at a voltage which is lower than the sum of their individual cut-in voltages. Therefore, sufficient diodes are connected in series such that the current leakage at the operating supply voltage remains negligible. Since N diodes in series are needed and the on-resistance of the clamp should be lower than the on-resistance of the input diodes, these diodes are sized at least n times larger than the input diodes. In this way, the supply clamp withstands higher ESD-stress than the input diodes ensuring that the latter is the limiting factor in the ESD-performance. The increased capacitance of the diode string is not a problem since it just adds to the on-chip decoupling capacitance and does not influence the RF performance.

5. Measurement results of a 1.57GHz LNA

5.1. RF measurement results

This LNA discussed in this section was designed for a portable GPS receiver [23]. It was implemented in a standard $0.25\mu m$ 4M1P CMOS process and occupies an area of $0.73mm^2$. For the RF measurements, the LNA is glued on a ceramic substrate and is wire bonded to 50Ω strip-lines. The substrate is then mounted in a Copper-Beryllium box, serving as a reference ground. The LNA is biased in two operating regimes drawing 4mA and 6mA from a 1.5V supply. The complete S-parameter has been measured together with the noise figure and IIP3. The main results are summarized in Table 2 and discussed more thoroughly in [23]. A photograph of the IC is shown in
Figure 7.

Table 2: Summary of the measured RF-performance of the LNA

	Bias condition A	*Bias condition B*
Power consumption	6mW	9mW
Current consumption	4mA	6mA
Supply voltage	1.5V	1.5V
S_{11} @ 1.57GHz	-12dB	-13dB
S_{21} @ 1.57GHz	15.5dB	16.5dB
S_{12}	< -30dB	<-30dB
S_{22} @ 1.57GHz	-14dB	-14dB
NF @ 1.57GHz	1.5dB	1.3dB
IIP3	-7dBm	-5dBm

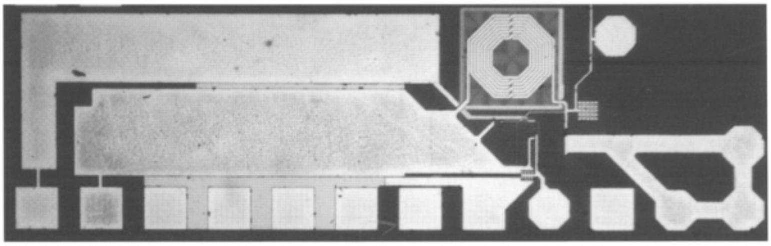


Figure 7: Micrograph of the LNA.

5.2. ESD-protection results

Figure *8* represents the measured Transmission Line Pulse (TLP) characteristics of the LNA circuit for three different stress combinations: INPUT to V_{DD}, INPUT to V_{SS} and V_{DD} to V_{SS}. The plot represents both the high current device IV characteristics $I_{ESD}=f(V_{dev})$ and the leakage current between the stressed pins, $I_{leakage}=f(I_{ESD})$. The jump in $I_{leakage}$ indicates the ESD failure threshold level. The TLP reverse breakdown characteristic of D_2 is also shown.

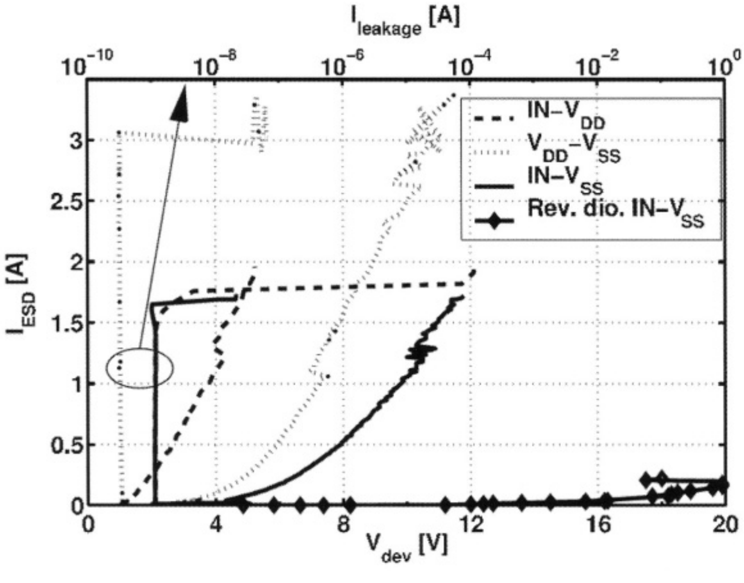


Figure 8: TLP results for three stress combinations on the nodes.

The results show that the TLP ESD robustness of the input diodes is of about 1.67A which corresponds to 2.5kV Human Body Model (HBM) ESD stress and the diode resistance is in the order of 2Ω. The ESD robustness of the D_3-D_7 diode stack is ~3.25A, corresponding to ~4.9kV HBM stress and a total resistance also in the order of 2Ω. The correlation of the TLP results to the HBM performance is for a first order evaluation only and is done by transforming the TLP I_{t2} current through the 1.5kOhm HBM model resistance. Table 3 represents the ESD thresholds, achieved from the actual separate on wafer HBM testing.

Table 3: Comparison between the achieved TLP and HBM thresholds for the major pin stress combinations.

Stress Polarity	*TLP: I_{t2} [A]*	*HBM V_{HBM}[V]*
IN^+ V_{DD}^-	1.67	3.2k
IN^+ V_{SS}^-	1.62	3.2k
V_{DD}^+ V_{SS}^-	3.2	>4k
V_{SS}^+ V_{DD}^-	4.2	>4k

The worst case ESD stress combination for the circuit is when it is stressed between the input and ground nodes. In this case, the ESD current flows from the input pad through D_1, the V_{dd} bus, the D_3-D_7 stack and the V_{ss} bus to the V_{ss} output pad. The overall ESD robustness in this case is determined by the lowest of the ESD thresholds of the different components in the current path: D_1, D_3-D_7, the reverse breakdown of D_2 and the ESD robustness of the input gate of the LNA. The measured ESD TLP threshold in this case was 1.62A. The results for the other major stress combination are represented in Table 3. The other possible combinations can be represented by these results as well.

6. LNA with inductive ESD-protection

The discussion in section 2 has made clear that both the noise figure and the power gain of the LNA degrade with the addition of the parasitic capacitance of a standard ESD-protection network. This effect becomes even more pronounced as the operation frequency increases. It becomes clear that a new ESD protection topology is required for these high frequency applications. This section presents a different input protection methodology for high frequency, CMOS LNAs. An on-chip inductor is employed to drain off the hazardous ESD-charge while in the mean time tuning out the harmful parasitic input capacitance. This technique will be demonstrated with the design of a 5.2GHz LNA attaining a noise figure of 2.1dB, and a power gain of 17dB in standard 0.25μm CMOS. The LNA input protection level complies with the 2kV HBM standard.

6.1. LNA with inductive ESD-protection

For >5GHz LNAs, the ESD-currents have a low frequency compared to the RF-signal. Therefore the ESD-protection network should act as a low-pass filter for the input current while the RF-signals should see a high-pass filter towards the LNA input. This can be realized by using a simple LC-filter (L_{ESD} and C_C in Figure 9).

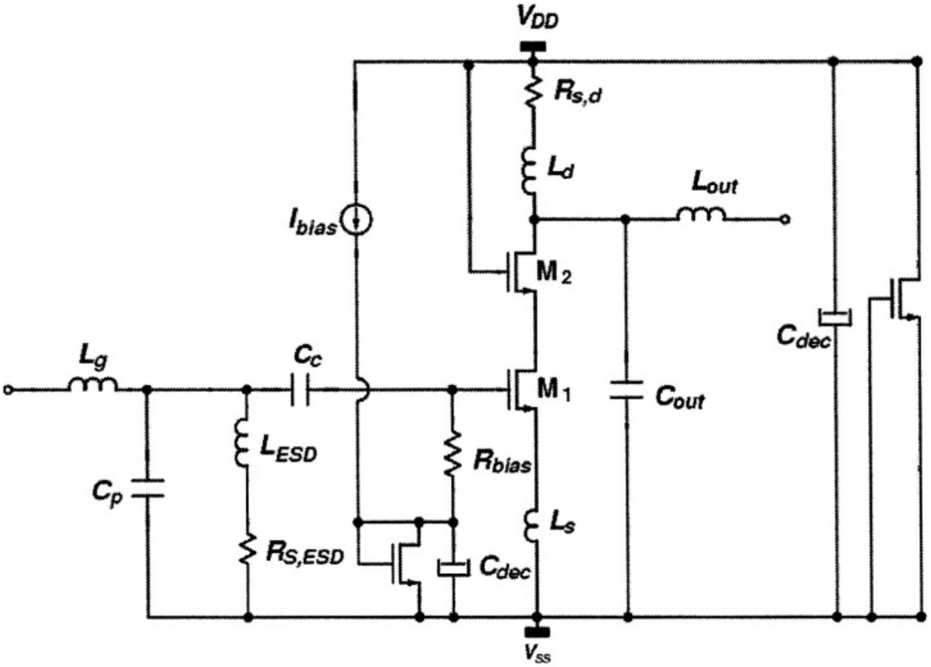


Figure 9: Schematic of an LNA with inductor as ESD-protection.

In Figure 9, an LNA is proposed, employing an on-chip inductor, L_{ESD}, to provide ESD-protection [22]. The inductor is depicted with its finite series resistance $R_{s,ESD}$ since it is to be realized on chip. In fact, this series resistance equals the on-resistance of the protection in case of an ESD event. Hence this resistance needs to be minimized from an ESD standpoint. At the relatively low signal frequencies of an ESD-pulse, the inductor conducts this charge with a minimum voltage drop over $R_{s,ESD}$.. At RF frequency, the inductor is designed to resonate with the input capacitance in order to avoid loss of signal. Where a normal ESD-protection structure would add parasitic capacitance to the input, this inductor tunes out all parasitic capacitance present at the input. It can be seen from the aforementioned equations that, ideally, this even increases the RF-performance. The value of R_{eq} is now simply 50Ω; the power gain increases and the noise figure decreases.

For very high frequencies, the inductor is very small and has only a small series resistance. Since the inductor is sized to tune out the exact amount of parasitic capacitance at the input, the only remaining parasitic at the input is the equivalent parallel resistance, $R_{p,ESD}$, of the resonant ESD-protection network. It is given by

$$R_{p,ESD} = \frac{\omega_0^2 L_{ESD}}{R_{s,ESD}} \tag{7}$$

This resistance increases the noise figure. From an RF performance point of view, it is clear that this resistance should be as high as possible and hence $R_{s,ESD}$ should be minimized (both from RF and ESD standpoint). Its influence would stay minor as long as

$$R_{p,ESD} >> \frac{1}{\omega_0^2 C_{gs}^2 R_s}, \tag{8}$$

where $R_s=50\Omega$. This implies that a decrease in the noise contribution of $R_{p,ESD}$ requires a larger input device, which means a larger power consumption. Moreover, a larger input transistor also implies a larger g_m (for the same V_{gs}-V_T), which increases the noise figure contribution of the drain noise current (1). Hence, the sizing of the input transistor is based on a trade-off between the different noise contributions and the power consumption. Generally, the contribution of $R_{p,ESD}$ to the noise figure will be dominant as the power consumption is a severe constraint.

The inductor is implemented as an octagonal coil and not as a rectangular coil. Therefore all corners have an angle of 135 degrees instead of 90 degrees. This is beneficial for the large ESD currents through the coil which are then spread more evenly over the coil width and avoid hot spots. The V_{DD} protection clamp is realized with a grounded gate NMOS (GGNMOS). It can sink a current of about 10mA/μm. In the design of this transistor, a protection level of 4kV HBM was aimed at. Since the maximum ESD current for an HBM-pulse is about 0.7A/kV the transistor should take 2.8A giving a width of 280μm which was rounded to 300μm to include some safety margin. The GGNMOS is implemented with several fingers. Drain ballasting is used to ensure a uniform triggering of all fingers.

At low operating frequencies, the required inductance for tuning out the input capacitance is very large since it is inversely proportional to the square of the operating frequency. This would also imply a large series resistance for the inductor. This cannot be tolerated from an ESD-performance perspective. One way to avoid this high on-resistance, is to deliberately increase the input capacitance. This can be realized by placing another parallel ESD-path increasing the ESD-immunity even further. The simplest solution is to place two reverse biased diodes in parallel with the inductor: one connected to V_{DD} and one to ground. These diodes will not only reduce the required inductance, they will also reduce the ESD-current flowing through the inductor, decreasing the on-resistance of the ESD-path further.

6.2. Results for a 5.2 GHz LNA

Simulations were done as in section 3. Figure 10a, b and c compare the LNA, completed with the proposed ESD-protection, with the same LNA, optimized without any ESD-protection. Both circuits are designed to consume 9mW from a 1.5V supply. Figure 10a depicts the LNA power gain (S_{21}). The ESD-protected LNA features a power gain of 17dB which is 4dB higher than the LNA without protection. This is due to the fact that all parasitic input capacitance, C_p, has now been tuned out with the ESD-protection inductor. This capacitance otherwise increases the equivalent source resistance, R_{eq}, seen by the gate of M_1, which yields a lower output current, as was shown earlier.

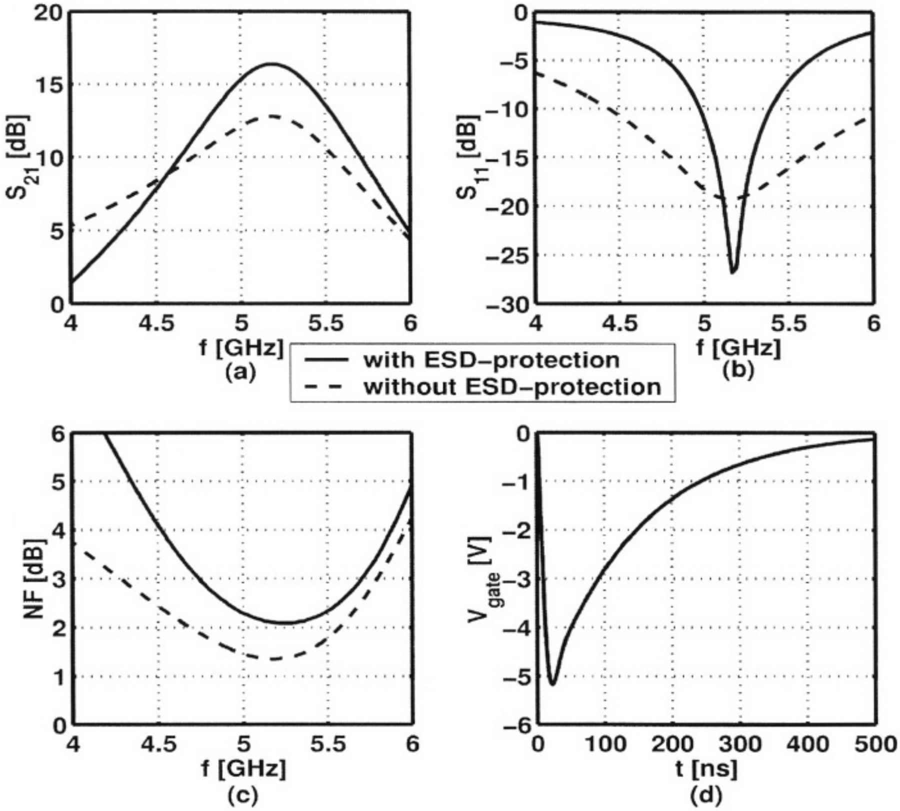


Figure 10 (a) Power gain of the LNA.
(b) Input reflection of the LNA.
(c) Noise Figure of the LNA.
(d) Input gate voltage under 2kV HBM stress.

Figure 10b shows the input reflection (S_{11}). It stays below –10dB in a 400MHz band for the LNA with the input protection inductor. For the LNA without the protection, the reduction in input reflection bandwidth is due to the introduction of the parallel resonance peak. Figure 10c plots the noise figure of both LNAs. Adding the ESD-protection increases the noise figure from 1.4dB to 2.1dB. The reason for the increase is the contribution of the $R_{p,ESD}$ of the tuned input tank. Figure 10d demonstrates the behavior of the voltage at the gate of M_1 when applying an HBM pulse of -2kV between the input pad and the V_{DD}. The maximum absolute value of the voltage over the gate oxide stays below 7V eliminating the danger for oxide breakdown.

7. Π-network implemented as ESD-protection structure

In section 2 we have discussed the influence of the equivalent input resistance on the LNA performance. It has been shown that C_p and L_g act as an impedance transformer where C_p incorporates the parasitic capacitance of the ESD-protection. The idea developed in this section is to use a CLC, i.e. a π-matching network (Figure 11), instead of a simple LC network. This network has the advantage that an extra degree of freedom is created, i.e. the network Q is decoupled from the impedance transformation ratio.

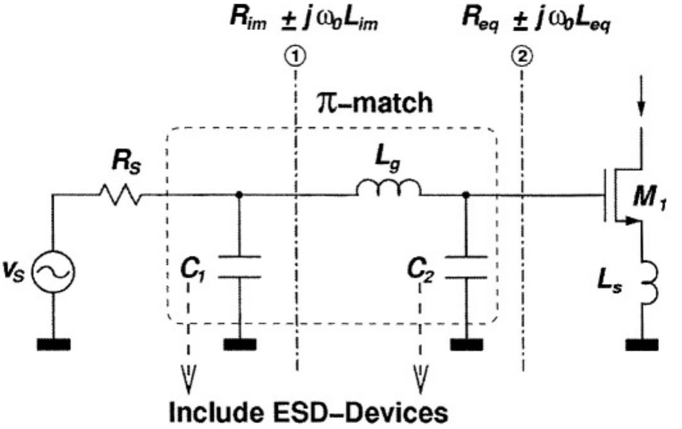


Figure 11: LNA input structure with π-matching network.

From an ESD-standpoint this network can be implemented by 2 ESD devices with an inductor in between. This network shows a remarkable resemblance with the standard input ESD-protection structure for digital CMOS mentioned earlier. Only here the resistance has been replaced by an on-chip inductor. The first ESD device is usually the largest and should take most of the ESD-current. The resistance limits the ESD-current flowing towards the input of the digital circuitry. The second device clamps the input voltage to tolerable level. The inductor does not work as well as a resistor since the impedance at the ESD-frequencies is lower. But it generates much less noise. The only noise stems from the series resistance which is in the order of 0.5 to 2Ω/nH. The inductance in its turn depends on the operation frequency, the network Q and the transformation ratio.

8. The common gate LNA

In the CG amplifier the signal is fed to the source of the input transistor rather than the gate. The impedance seen at the input of the LNA (neglecting the input capacitance) is then,

$$R_{in} = \frac{1}{g_m + g_{mb}} \left(\frac{r_{ds} + R_{load}}{r_{ds}} \right) \tag{9}$$

It can be set to equal the impedance of the source, R_s, usually 50Ω. The power gain of the LNA is now simply given by,

$$G_t = \frac{R_{load}}{4 \cdot R_s}. \tag{10}$$

The noise factor of the LNA is approximated by,

$$F = 1 + \left(\frac{\gamma}{\alpha} \right) \left(\frac{g_m}{g_m + g_{mb}} \right) \left(\frac{1}{R_{load} \cdot \lambda \cdot I_d} \right). \tag{11}$$

It is seen that the noise figure can theoretically be made arbitrary low by simply pumping enough current through the transistor. The IIP3 of the CG LNA is approximated by,

$$IIP3[\mathrm{dBm}] \approx -10 + 10\,log\left(\frac{4}{3} \frac{V_{gt}(1 + \Theta V_{gt})^2 (2 + \Theta V_{gt})}{\Theta} \right). \tag{12}$$

In case the input capacitance is too large, it can be tuned out with a parallel inductor; equations (9) to (12) remain valid.

The performance of the CG LNA with respect to noise figure and gain is lower than that of the CS LNA at low frequencies. However for the CS LNA the excess noise is for the most part proportional to the square of the operation frequency and the power gain is inversely proportional to the square of the frequency. Consequently, the performance rapidly declines with increasing frequency. And at higher frequency the CG LNA performs better then its CS counterpart [24]. To demonstrate this a figure of merit is given in (13).

$$FOM = G[\mathrm{dB}] - 10 \cdot log(F - 1) + IIP3[\mathrm{dBm}] \tag{13}$$

The FOM in (13) is plotted vs. frequency in Figure 12a for different power budgets. It can be seen that at a frequency in the order of 12GHz, it becomes more interesting to use a CG LNA. Figure 12b plots the noise figure for both types of amplifier as a function of operating frequency for different power budget. Also here there is a crossing frequency and it is in the order of 8 to 10 GHz The exact crossing frequency in both plots depends on the power consumption but even more so on the used technology. This plot was made for a 0.25μm technology. Generally the crossing frequency will increase for smaller technologies due to the scaling of the cut-off frequency (ω_T).

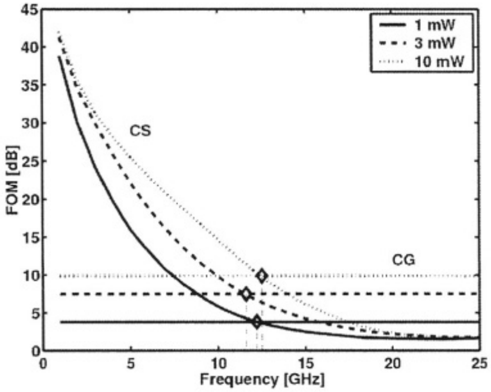


Figure 12a: FOM for CS and CG LNA as a function of frequency.

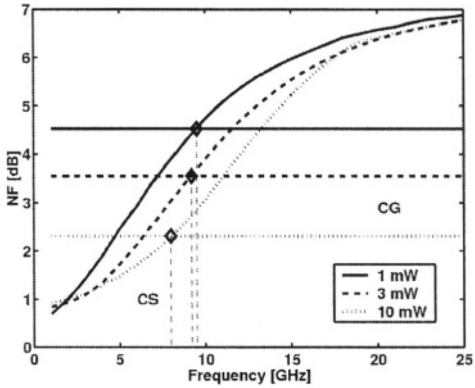


Figure 12b: NF for CS and CG LNA as a function of frequency.

With regards to ESD, the CG LNA is less sensitive than the CS LNA since the input is not at the gate but at the source of the transistor. As a consequence the input already has the inherent source-bulk junction which is able to carry the charges for a negative pulse vs. ground. An additional protection can be provided by the parallel inductor, used to tune out the excess input capacitance. This is basically the same strategy as in section 6. Extra ESD-devices can be added, increasing the input capacitance and lowering the inductance needed to tune it out. Also the Π-network protection strategy can be used since ideally it is a lossless network and only leads to a transformation of the equivalent input impedance. The discussion is similar to section 7.

9. Conclusion

The two most important CMOS LNA topologies have been discussed. A rigorous RF-ESD co-design methodology has been explained for high performance CMOS LNAs. This methodology has been discussed by means of the inductively degenerated common source LNA. Results of this design strategy were demonstrated with the measurements of an LNA for a portable GPS receiver. This paper further discussed two different ESD-protection strategies. The first is based on the use of an inductor as a signal splitter, the second on the use of a Π-network incorporating the ESD-protection. For a glimpse of the future, a comparison was made between the performance of the CS and CG LNA and some promising ESD-protection methodologies for the latter have been discussed. This work shows that, even in a standard submicron CMOS technology, a high RF-performance may be combined with a good level of ESD-protection satisfying the industrial specification of 2kV HBM.

References

[1] T.J. Maloney and S. Dabral, “Novel Clamp Circuits for IC Power Supply Protection,” in Proc. EOS/ESD Symp., 1995, p.1.

[2] A. N. Karanicolas, “ A 2.7V 900MHz CMOS LNA and Mixer,” ISSCC Digest of Technical Papers, pp. 50-51, 416, Feb. 1996.

[3] A. Rofougaran et al., “A 1GHz CMOS RF Front-End IC for a Direct-Conversion Wireless Receiver,” in IEEE Journal of Solid-State Circuits, vol. 31, no.7, pp. 880-889, July 1996.

[4] D. K. Shaeffer and T. H. Lee, “A 1.5V 1.5GHz CMOS Low Noise Amplifier,” in IEEE Journal of Solid-State Circuits, vol. 32, no.5, pp. 745-759, May 1997.

[5] J. Janssens, M. Steyaert and H.Miyakawa, “A 2.7 Volt CMOS Broadband Low Noise Amplifier,” Digest of Technical Papers, Symposium on VLSI circuits, pp. 87-88, Kyoto, June 1997.

[6] A. R. Shahani et al., “ A 12-mW Wide Dynamic Range CMOS Front-End for a Portable GPS Receiver,” IEEE J. Solid-State Circuits, vol.32, no. 12, pp. 2061-2070, Dec. 1997.

[7] J. Zhou and D. J. Allstot, “ A Fully Integrated CMOS 900 MHz LNA utilizing Monolithic Transformers,” ISSCC Digest of Technical Papers, pp. 132-133, Feb. 1998.

[8] C. P. Yue and S. S. Wong, “On-chip Spiral Inductors with Patterned Ground Shields for Si-Based RF IC’s,” in IEEE Journal of Solid-State Circuits, vol. 33, no.5, pp. 743-752, May 1998.

[9] F. Stubbe et al., “ A CMOS RF-Receiver Front-End for 1 GHz Applications,” Symp. on VLSI Circuits Dig., 1998.

[10] G. Hayashi et al., " A 9 mW 900MHz CMOS LNA with Mesh Arrayed MOSFETs," Symp. on VLSI Circuits Dig., 1998

[11] Q. Huang et al., " Broadband, 0.25µm CMOS LNAs with Sub-2dB NF for GSM Applications," Proc. IEEE Custom Integrated Circuits Conference, pp. 67- 70, 1998

[12] J. Janssens and M. Steyaert, "MOS noise performance under impedance matching constraints," Electronics Letters, vol. 35, no. 15, pp. 1278-1280, July 1999.

[13] A.Amerasekera et al.., "Analysis of snapback behaviour on the ESD capability of sub-20um NMOS," in Proc. IRPS 99 , pp. 159-166, 1999.

[14] B. A. Floyd et al., "A 900MHz 0.8µm CMOS Low Noise Amplifier with 1.2dB Noise Figure," in Proc. IEEE Custom Integrated Circuits Conference, pp. 661-664, 1999.

[15] G. Gramegna et al., " Ultra-Wide Dynamic Range 1.75 dB Noise-Figure, 900MHz CMOS LNA," ISSCC Digest of Technical Papers, pp. 380-381, Feb. 2000.

[16] S. Dabral and T. J. Maloney, "Basic ESD and I/O Design" published by Wiley Interscience

[17] P. Leroux, J. Janssens and M. Steyaert, "A 0.8dB NF ESD-protected 9mW CMOS LNA", in ISSCC, Digest of Technical Papers, (San Francisco), pp. 410-411, Feb. 2001.

[18] G. Gramegna, M. Paparo, P. G. Erratico, and P. De Vita, "A sub-1-dB NF ±2.3-kV ESD-Protected 900-MHz CMOS LNA", in IEEE Journal of Solid-State Circuits, vol. 36, no.7, pp. 1010-1017, July 2001.

[19] C.Richier et al., "Investigation on Different ESD protection strategies devoted to 3.3V RF applications in a 0.18 CMOS process' Proc. EOS/ESD Symp. 2000, pp.251-259, September 2001.

[20] M. Radhakrishnan et al., "ESD reliability issues in RF CMOS circuits", in Proc. IWPSD 2001, International Workshop on the Physics of Semiconductor Devices, December 2001.

[21] F. Bruccoleri, E.A.M. Klumperink and B. Nauta, "Noise Cancelling in Wideband CMOS LNAs," in ISSCC, Digest of Technical Papers, (San Francisco), Feb. 2002.

[22] P. Leroux, M. Steyaert, "A New ESD Protection Topology for High Frequency CMOS Low Noise Amplifiers", in Proc. EMCE, pp. 129-133, September 2002.

[23] P. Leroux, V. Vassilev, M. Steyaert, G. Groeseneken, "A 1.3dB NF CMOS LNA for GPS with 3kV HBM ESD-protection", in Proc. ESSCIRC, pp. 335-338, September 2002.

[24] X. Guan and A. Hajimiri, "A 24GHz CMOS Front-end," in Proc. ESSCIRC, pp. 155-158, September 2002.

IMPROVEMENT OF SYSTEM ROBUSTNESS THROUGH EMC OPTIMIZATION

Bernd Deutschmann
austriamicrosystems AG
Schloss Premstätten
A-8141 Unterpremstätten
Tel.+43 3136 500-5971
bernd.deutschmann@austriamicrosystems.com

Abstract

The electromagnetic compatibility became a major concern for integrated circuits during the last years. As the operating frequency has increased and the supply voltage has become lower, the electromagnetic emission as well as the susceptibility of integrated circuits increased tremendously. Generally speaking, electromagnetic compatibility is defined as the ability of an electrical system to work properly in its electromagnetic environment without unduly interfering with this environment. This is the reason why electromagnetic compatibility is of importance to all of us. Consider, for example, the omnipresent electromagnetic environment in our motor vehicles generated by devices like ABS braking systems, airbag sensors, or the motor management. Without a proper design, electromagnetic compatibility of integrated circuits will become a limiting factor for the performance of every advanced electronic system.

In this paper, the new standards to characterize the electromagnetic compatibility of integrated circuits will be presented. These standards provide measurement methods to characterize the electromagnetic emission and the immunity of integrated circuits. It will be shown how two measurement methods have been used to find the source of the unwanted electromagnetic emission of an integrated circuit.

A. van Roermund et al. (eds.), Analog Circuit Design, 227-242.
© 2003 *Kluwer Academic Publishers. Printed in the Netherlands.*

1. Introduction

Today the electromagnetic compatibility (EMC) of integrated circuits (ICs) has emerged as a principal theme. The need for low emission of electromagnetic energy as well as a high-level of immunity has caused EMC to be as important as performance and functionality. The manufacturers of electronic devices and systems are forced to develop their products according to EMC requirements. However, the semiconductor manufacturers will also have to consider electromagnetic emission and susceptibility of their products. A design engineer can no longer assume that his system in which modern ICs are implemented will meet the compulsory limits and performs compatible with other electronic systems. In the past, the major concerns of the semiconductor industry were cost, area, performance, and reliability; EMC was mostly of no or secondary importance.

Although an IC usually does not radiate significantly, the IC is often seen as the source of overall radiated electromagnetic energy of an electronic system. Many problems originate from the fact that most of the digital signals are switched as fast as possible, resulting in an increasing amount of high frequency spectral content.

Figure 1 shows these technology trends by means of ASICs. As can be seen the operating frequency, as well as the number of transistors being integrated on a die, has increased remarkably during the last years. On the other hand, the supply voltage and the channel length of the transistors have decreased in order to be able to handle problems caused by size and power dissipation, which for example makes ICs more vulnerable.

Especially the susceptibility to radio frequency signals (radio frequency interference RFI), coming from mobile phones as well as high speed wireless network systems has gained more and more significance. In general, ICs arc extremely susceptible to these signals and may react in an unwanted manner. This often causes failures of the designed function and may be safety critical.

For this reason implementing EMC measures on chip-level is absolutely necessary to meet today’s demands. Especially for automotive applications, EMC characterization on chip-level became a significant attribute of quality and thus a decisive criterion for a customer’s choice of components.

Today’s improvements concerning chip-level performance, speed, and technology sound wonderful. However, problems become evident when you imagine the consequence of agglomerating many ICs into a system-level product.

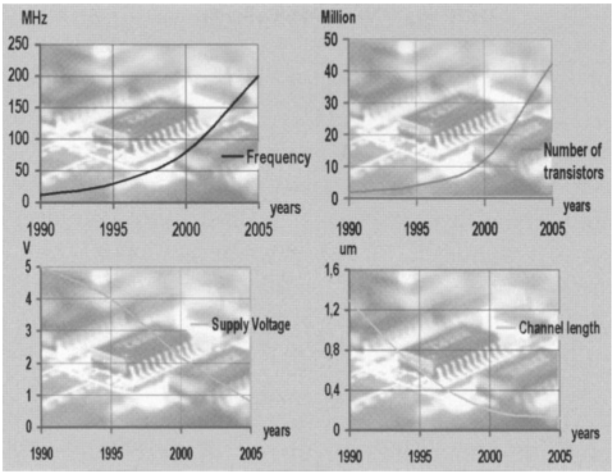


Fig. 1. Technology trend (ASICs).

The emission of electromagnetic energy will increase at an exponential rate, whereas the immunity will decrease. Mathematically, unless the EMC performance increases with the same exponential rate as the ICs improvements, the EMC will eventually become a serious impediment to overall system performance. That is happening right now.

This does not mean that all the EMC experts have been resting on their laurels, but in fact we can count on the fingers of one hand the number of big, widespread advances in EMC on chip-level within the last years. With placing the VDD and GND pins close to each other, implementing on-chip decoupling, controlling the slew rate, making analog parts more robust being some of the prominent advances. So far, these advances were sufficient to keep up with the chip industry, but in the future further progress will be necessary.

To fully realize the benefits of advanced ICs, here is what will be required: “Better knowledge of how the ground bounce and consequently common mode currents affect the electromagnetic emissions, better package models for simulation, better tools for EMC management on chip-level, reduction of signal-integrity problems, higher immunity of analog building- and I/O-blocks to RF interference, better control over radiated emissions, better packages with less parasitic elements, advanced control of the rise and fall times of the output signals (moderate slew rate), and of course more design engineers who understand the generation of ICs emissions and how to improve their immunity.”

If we follow all these points, we can expect to keep up with Moore’s Law for a long time.

1 Characterizing the electromagnetic compatibility of ICs

Worldwide there are about 200 standards to characterize the electromagnetic compatibility of electronic devices and systems. As the EMC characterization on chip-level is becoming more and more important we will also need well-defined standards and emission reference levels for ICs used within these electronic systems. The subcommittee 47A (Integrated Circuits) of the IEC[1] is working towards two standards for the characterization of the electromagnetic emission (IEC 61967 [1]) and the electromagnetic immunity (IEC 62132 [2]) of ICs. Although the measurement methods described in these two standards will never be able to completely eliminate EMC measurements on system-level, a system design engineer will have the ability to identify the main sources of the emissions and the parts having the weakest immunity in his application.

Under the general title "Integrated circuits – Measurement of electromagnetic emissions 150kHz to 1GHz" the current version of the IEC 61967 standard consists of six parts: First, there is the general conditions and definitions part and the five following parts describe different measuring methods to characterize the conducted and radiated electromagnetic emissions of ICs. In ascending order there are the TEM cell method and the Surface Scan method for the measurement of the radiated emissions and the **1Ω/150Ω** Direct Coupling method, the Workbench Faraday Cage method as well as the Magnetic Probe method are applied for the measurement of the conducted emissions.

At present the second standard "Integrated circuits - Measurement of electromagnetic immunity" (IEC 62132) consists of 5 parts. Again, there is a general conditions and definitions part and by now four different measuring methods to characterize the radiated and conducted immunity of an IC. There is again a TEM cell method for the measurement of the radiated immunity as well as a Bulk Current Injection method (BCI), a Direct Power Injection method (DPI), and again a Workbench Faraday Cage method for the measurement of the conducted immunity.

All measuring methods described in these two standards can be used as a basis for the specification of the emission and the immunity of ICs. All these methods have their advantages and limitations. Therefore, the designer of electronic applications, as well as the semiconductor manufacturer, should carefully select those measuring methods, which best fit their requirements. Although we are able to characterize the EMC of ICs on chip-level using these measuring methods, it will not be possible to directly compare between system- and chip-

---

[1] The IEC (International Electrotechnical Commission) is a worldwide organization for standardization comprising all national electrotechnical committees. The object of Ihe IEC is to promote international cooperation on all questions concerning standardization in the electrical and electronic fields. To this end and in addition to other activities the IEC publishes International Standards.

level measurements. The manufacturer of electronic applications will always have to perform EMC measurements of the whole system, even if the IC has passed the EMC requirements on chip-level.
Often the typical EMC measures like filters or shielding techniques are necessary to fulfill the EMC requirements of the electronic application. Where those EMC measures are realized (within the IC or for example on the printed circuit board) depends on cost aspects and feasibility considerations (space availability, etc.).
The most effective way of solving EMC problems is to pinpoint and to reduce the actual sources of the interference. A very important way to tackle EMC problems already on chip-level will be shown in the last chapter.

1.1 Emission measurements / the Emission level scheme

To determine the electromagnetic emission of the IC under test in the full spectrum range from 150kHz - 1GHz a so-called emission level scheme is used. Using this classification technique, comparisons between different ICs can be done in a standardized way. The emission level scheme illustrates the maximum emission level with two letters and one number always following the same sequence. The capital letter, which is first, represents the position of the horizontal line having a slope of 0dB/decade. Second is a number, which defines the position of the -20dB/decade slope. The third and small letter defines the position of the -40dB/decade slope. Assuming that one of the three slopes is not needed then the corresponding letter or number will be omitted. Figure 2 gives an example how this emission level scheme is used to describe the maximum emission of an IC. In this case the IC will be classified having a maximum emission level of “B2a”.

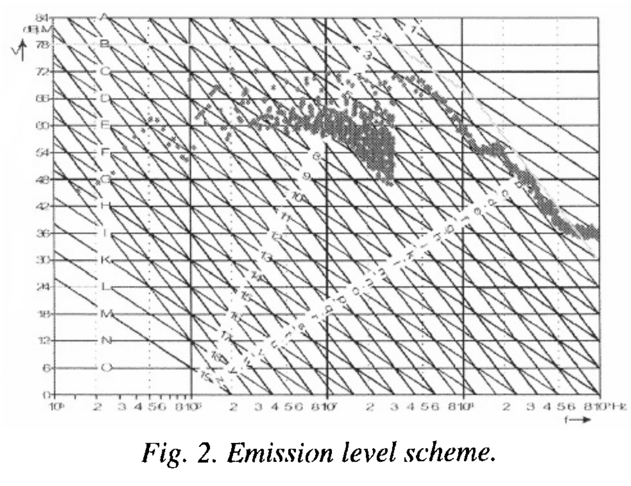


Fig. 2. Emission level scheme.

1.2 Immunity measurements / performance classes

The value of the immunity level depends on system and application specific parameters. To qualify ICs a simple measurement procedure is needed having a high degree of repeatability. The electromagnetic immunity of an IC can be characterized by conducted or radiated RF-disturbanccs. In both cases, the RF power needed to cause a malfunction of the IC under test is measured. Since the IEC 62132 standard does not suggest criteria to determine whether or not the IC under test is working properly, these criteria have to be defined by the user himself. In part 1 of the standard several performance classes are described (from A to D) to classify the malfunctions.

Class A: Normal performance within the specification limits during and after the exposure to a disturbance.
Class B: Temporary degradation or loss of function or performance, which is self-recoverable (e.g. when scanning further through the frequency range).
Class C: Temporary degradation or loss of function or performance, which requires operator intervention or system reset.
Class D: Degradation or loss of function which is not self recoverable due to damage of IC(s), or loss of data.

For the test continuous wave (CW) as well as amplitude-modulated (AM) signals should be used. Typically an 80% amplitude modulation of 1kHz is used. Any function inside the IC can be affected even if it is not connected to the pin at which the RF power is applied. So the IC shall be driven in an operation mode that all the different functions are used during the test. The necessary immunity threshold level, which often depends on the application of the IC and the pin, which is tested, should carefully be chosen to avoid over-testing of the IC. Each pin, which is expected to be submitted to RF interference, should be tested individually. Also multiple pin testing can be performed if several pins have identical circuitry and layout.
If the designer of an electronic system knows about the immunity of an IC, he can decide if he needs external protection circuitry on the printed circuit board (PCB) to satisfy the EMC criteria.

2 The different measuring methods to measure the electromagnetic emission and immunity at a glance [3]

2.1 IEC 61967-2 and IEC 62132-2 (TEM-Cell method)

The TEM cell method can be used to measure the electromagnetic radiation from an IC under test (IEC 61967-2 [4]), as well as to identify its immunity to electromagnetic fields (IEC 62132-2 [5]). In both cases the IC under test is not placed as usual in the TEM cell, as it is done when electronic devices and systems are tested, but it is mounted on a standardized IC test PCB, which has to be designed according to the standard. This test PCB is clamped to a mating port, which is cut in the top or bottom of a TEM or wideband TEM (GTEM) cell. The IC under test and the test PCB become a part of the cell wall with the IC under test inside the TEM cell.

This test PCB has to be designed in a way that all connecting leads within the cell are eliminated. All the connecting leads should be located on the backside of the board outside the cell. Only the IC is allowed to be inside the cell.

The TEM cell has two 50Ω ports. One of these ports is terminated with a 50Ω load. For the emission measurement the other 50Ω port of the TEM cell (or the single 50Ω port of the GTEM cell) is connected to the input of an EMI receiver or spectrum analyzer to measure the RF emissions emanating from the IC.

For the immunity measurement, the 50Ω port is connected to a RF generator, which is used to feed electromagnetic energy into the TEM cell. The electromagnetic field generated by the TEM cell has an impact on the IC and degrades its functionality. The amplitude of the RF generator is increased until the maximum power level is reached or the monitoring device notes a malfunction of the IC under test.

Figure 3 shows an example of the described TEM cell with a mounted PCB.

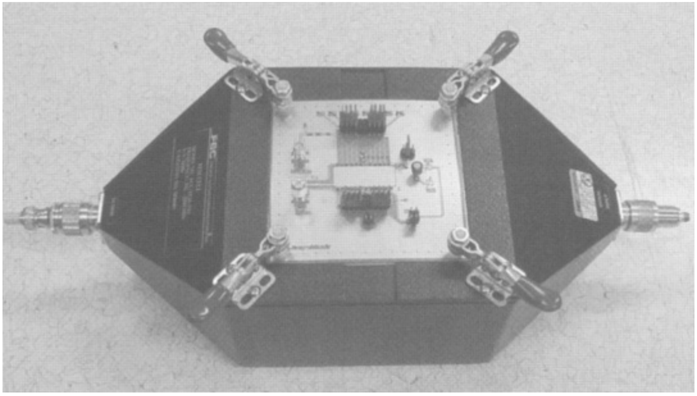


Fig. 3. TEM cell with mounted PCB.

2.2 IEC 61967-3: The surface scan method

This part of the IEC 61967 standard describes a method for evaluating the near electric (E) field and the near magnetic (H) field component at the surface of an IC [6]. This method can be used over the frequency range from 10MHz to 3GHz. To measure the distribution of these fields an E- or an H-field probe is mechanically moved over the surface of the IC, where the probe can be placed in a parallel or perpendicular plane to the IC surface. The measured data can be computer processed and the field strength at a specific scan frequency can be visualized in a colored representation. The resolution reachable with this method is closely linked to the precision of the mechanical probe positioning system and the size of the used probes. In the last chapter it will be shown how comparisons between different design steps can be done in order to show improvements in the design. This measurement method can be applied to any IC mounted on a PCB that is accessible to the probe. For comparison measurements between different chip architectures, a standardized test board as described in the standard shall be used. The electric and magnetic field scan over the surface of the IC yields information about the electromagnetic emission. This method provides a useful tool to locate areas on the die with high electromagnetic energy, which often act as the sources of the electromagnetic emission of an IC.

E- and H-field Probes:

To measure the electric field, the mechanical design of a micro sized E-field probe, which usually has a partial shield, is suggested in the IEC 61967-3 standard. For H-field measurements, this standard recommends the use of a single turn micro sized H-field probe; either stand-alone or generated as a PCB trace. Both probes can be fabricated from a 0.020inch semi-rigid coaxial cable. Figure 4 shows an example of an E- and an H-field probe, built by a coaxial cable.

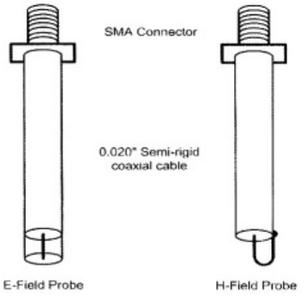


Fig. 4. E and H near-field probes.

2.3 IEC 61967-4 (1Ω/150Ω method)

The so-called 1Ω/150Ω method is described in the third part of the IEC 61967 standard [7], [8]. This method is used for the characterization of the conducted electromagnetic emission of the IC under test. First, there is the 1Ω method which is used to measure the RF sum current of a single ground pin. Next, we have the 150Ω method to measure RF voltage at an output pin.

In general, the conducted electromagnetic emissions are mainly caused by fast changes of the RF currents inside and outside of the IC. These currents are mainly distributed via the IC pins to the signal traces of the PCB. Often the layout of the traces on the PCB generates loops, which act as emitting antennas of the electromagnetic energy.

As most of the RF currents return via the ground or supply pin back to the IC, the ground return path is the best location to measure the conducted emissions using an RF current probe. This current probe consists of a 1Ω resistor, which is connected between the common ground pin of the IC and the ground signal trace on the PCB, together with a 49Ω matching resistor in order to connect a test receiver (spectrum analyzer) with an input-impedance of 50Ω. The voltage drop across the 1Ω resistor is measured by a test receiver and indicates the total emission of the IC under test.

With the 150Ω method the contribution of a single I/O pin or multiple I/O pins to the total emission of an IC can be evaluated. This can be done by measuring the RF voltage drop at a 150Ω impedance matching network. The 150Ω coupling network is built according to the characteristic impedance of lines in the cable harnesses in vehicles. For this reason the measurement method shall only be applied to those pins of the IC, which are connected to PCB traces or wiring harness longer than 10cm. A typical coupling network consists of a voltage divider built by the resistors R1, R2 together with the input resistance of the test receiver and a capacitor C1 to avoid excessive direct currents at the output.

The general test configuration for the 1Ω method (RF current probe) and the 150Ω method (Impedance Matching Network) is shown in figure 5. Due to the simple measurement setup, these two methods guarantee a high degree of repeatability and correlation of electromagnetic emission measurements.

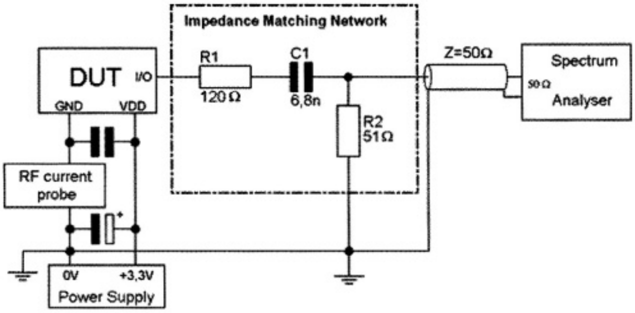


Fig. 5. General test configuration (1Ω/150Ω method).

2.4 IEC 61967-5 (Workbench Faraday Cage method)

The Workbench Faraday Cage (WBFC) method can be used for emissions as well as for immunity measurements of ICs [9], [10]. The PCB with the IC under test is placed inside a small Faraday cage to avoid interference to the operation of electronic equipment close to the cage during immunity tests and to make sure that only the emissions of the IC under test are measured.

All the functional connections, like power supplies, the auxiliary equipment, and I/Os are fed through filters having high common-mode impedances. These filters are mounted on the wall of the cage. The basic concept of this method was derived from the European standard EN 61000-4-6, which focuses on the immunity of electronic equipment to common mode conducted radio frequency interference.

The main electromagnetic emissions and interferences are caused by emissions and interferences collected by cables, which are connected to the IC pins. Depending on their length and geometry, these cables often build the receiving and transmitting antennas having an average radiation resistance value of about 150Ω.

Using the WBFC for emission measurements the common-mode impedances of these antennas are replaced by a 150Ω resistor. The voltage drop across this resistor is used to characterize the RF emission of the IC under test.

For immunity measurements, the antennas are replaced by an interference source and electromagnetic energy is fed into the faraday cage while the IC under test is monitored.

2.5 IEC 61967-6 (Magnetic probe method)

This part of the standard specifies another method to evaluate the RF emissions of an IC [11].

In this case the RF currents of the IC under test are estimated using a small magnetic probe, which is used to measure the magnetic field at a specified height over an I/O or power supply strip conductor on a standardized test board. The measuring frequency range is specified from 150kHz to 1GHz and higher. The detailed structure and the recommended dimensions of the magnetic probe can be taken from the standard. A triplate-structured microstrip line consisting of a three-layer PCB is used for this probe.

In figure 6 a principle scheme of the standardized test setup to measure the magnetic field strength over an output line is shown.

A high degree of repeatability is provided by this test method if the magnetic probe is placed accurately.

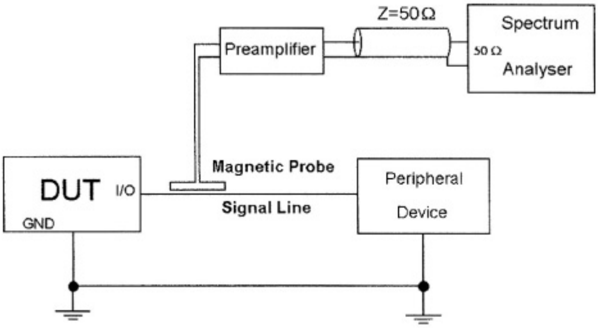


Fig. 6. Setup for emission testing (magnetic probe).

2.6 IEC 62132-3 (Bulk current injection method, BCI)

The Bulk Current Injection method (BCI), which is described in IEC 62132-3, is used to characterize the immunity of an IC to conducted electromagnetic disturbances [12]. With an inductive current injection probe, which is driven from the output of an RF signal source, energy is coupled into wires which are connected directly to the IC under test. The measuring test setup works properly within the frequency range of 10kHz to 400MHz. Beginning with the lowest frequency, the RF current is increased until a failure of the IC under test occurs or the maximum specified test severity level is reached. In the case of a failure, the interference amplitude determines the susceptibility threshold at this frequency. Figure 7 shows an example of a setup of the BCI measuring method.

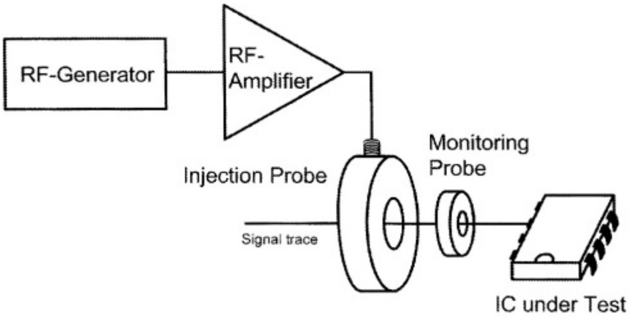


Fig. 7. Setup of the Bulk Current Injection method (BCI).

2.7 IEC 62132-4 (Direct RF power injection method, DPI)

The Direct Power Injection method (DPI) IEC 62132-4 is again a measuring method to determine the immunity of an IC under test to conducted RF disturbances [13]. The RF power is directly applied to a single pin or to a group of pins having the same circuitry and layout. A capacitor shall be used, which performs as a DC block to avoid supplying DC current into the output of the RF signal generator and the amplifier, together with a resistor to simulate a load of a real application. To avoid RF interference to the power supply of the IC under test, decoupling networks having high RF impedances are used.

3 Using the IEC 61967 standard to improve an automotive ASIC

Especially in the automotive business, the electromagnetic emission of ICs became a major concern. Therefore semiconductor manufacturers are trying to reduce the electromagnetic emission of their ICs by applying several EMC features on chip-level like on–chip decoupling, slew rate controlled output drivers, or filters.

The following example shows how two of the measuring methods described in the IEC 61967 standard were used to measure the total electromagnetic emission of an automotive ASIC and how the source of the high electromagnetic emission on the IC could be detected. With this information the designer is given hints which area of the IC is responsible for the high electromagnetic emission. He has now the ability to redesign this designated area in order to reduce the total electromagnetic emission of the IC.

First a TEM cell measurement of the radiated electromagnetic emission in the frequency range of 150kH/ up to 1GHz was performed to find the frequencies with too high emissions. The result of this measurement is shown in figure 8. As can be seen, high electromagnetic emissions can be found in the frequency range of about 150MHz.

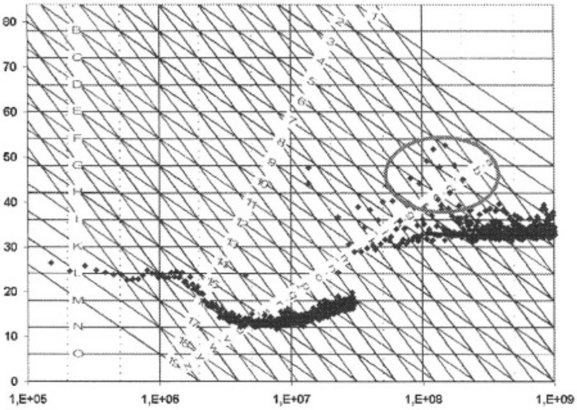


Fig. 8. Result of the TEM cell measurement.

Next, the surface scan method was applied in a chosen bandwidth around the frequency of 150MHz to pinpoint the source of the electromagnetic energy. In this case only the magnetic field distribution at the surface of the die was measured. The result of the surface scan is shown in figure 9. The area of highest electromagnetic emission was found in the upper right corner. Speaking with the designer, the circuit parts, which are responsible for the high emission of the IC, were detected immediately.

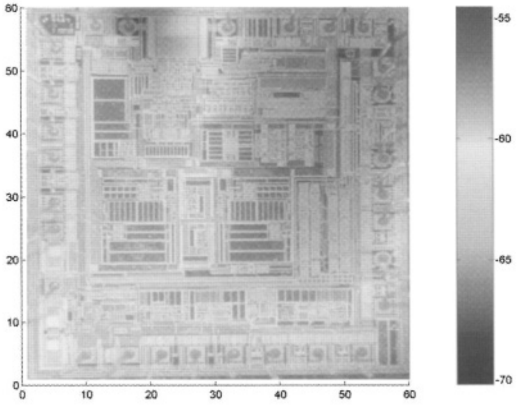


Fig. 9. Magnetic field at the surface of the die (result of the surface scan method).

From this investigation a redesign of these circuit parts was performed in order to reduce the high electromagnetic emissions in the 150Mhz range.
The surface scan method was again applied for the redesigned ASIC to verify that the redesign was successful. The result of the surface scan is shown in figure 10.

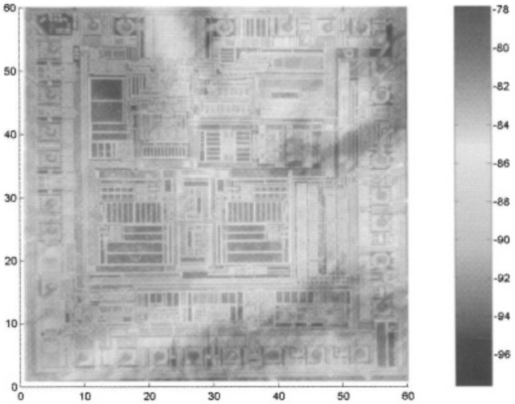


Fig. 10. Magnetic field at the surface of the die (result of the surface scan method of the redesigned ASIC).

For the comparison of figure 9 and 10 it is important to know, that the scale of the right column in figure 9 has the range –70 to –55dB and in figure 10 the range is –98 to –78dB. So the surface scan from the redesign shows that the maximum electromagnetic emission of the ASIC was reduced tremendously. Remember from the surface scan result of the original design that the area of high electromagnetic emission was found in the upper right corner. Now this area shows very low emissions. The redesign has not failed.

To check and to prove that the redesign did not affect other areas, which might now be responsible for higher electromagnetic emissions in different frequency ranges, a TEM cell measurement of the redesigned IC was performed again. The result is shown in figure 11.

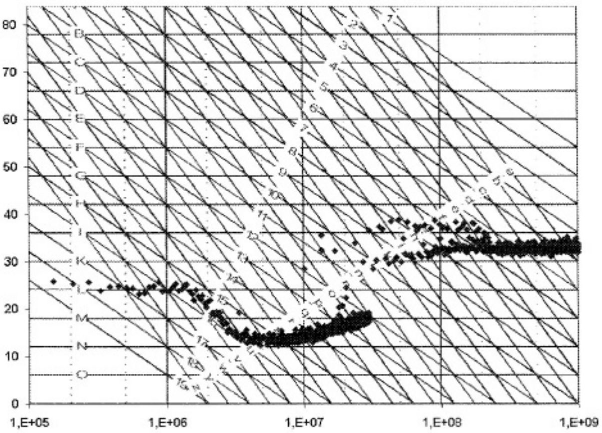


Fig. 11. Result of the TEM cell measurement of the redesigned ASIC.

Comparing figure 8 and figure 11 again shows that the high unwanted electromagnetic emission especially in the frequency range of 150Mhz could really be reduced.

4 Conclusion

This paper has focused on the characterization of the electromagnetic compatibility of integrated circuits. The main measurement methods for emission and immunity, which are proposed in two new standards, have been explained. The general conditions of the emission and the immunity measurements have been described.

Finally, it has been shown how two of the emission measurement methods have successfully been used to improve the electromagnetic emission of an automotive ASIC.

References

[1] IEC 61967-1: “Integrated circuits -Measurement of electromagnetic emissions, 150kHz to 1GHz - Part 1: General conditions and definitions”, 47A/632/FDIS

[2] IEC 62132-1: “Integrated circuits - Measurement of electromagnetic immunity, 150 kHz to 1GHz - Part 1: General and definitions”, 47A/618/CD

[3] B. Deutschmann, G. Winkler, R. Jungreithmair, ,,EMV-Charaktcrisierung von ICs” Test Kompendium 2003, pp 35-37, ISBN 3-934698-08-5

[4] IEC 61967-2: “Integrated circuits -Measurement of electromagnetic emissions, 150kHz to 1GHz - Part 2: TEM-cell method and wideband TEM-cell method (150kHz to 8GHz)”, 47A/619/NP, New Work Item Proposal, Date of proposal: Jul. 2001

[5] IEC 62132-2: “Integrated circuits - Measurement of radiated immunity - Part 2: TEM cell method” (to be proposed)”

[6] IEC 61967-3: “Integrated circuits - Measurement of electromagnetic emissions, 150kHz to 1GHz – Part 3: Measurement of radiated emissions, surface scan method (10kHz to 3GHz)”, 47A/620/NP, New Work Item Proposal, Date of proposal: Jul. 2001

[7] IEC 61967-4, “Integrated Circuits, Measurement of Electromagnetic Emissions, 150kHz to 1GHz Measurement of Conducted Emission, 1Ohm/150Ohm Method”, 47A/606/CDV, 2001-04-13

[8] B. Deutschmann, G. Winkler, R. Jungreithmair, ,,Measuring the Electromagnetic Emissions of Integrated Circuits with IEC 61967-4 (The Measuring Method and its Weaknesses)”, IEEE International Symposium on Electromagnetic Compatibility, ISBN 0-7803-7264-6/02, August 2002, pp 407-412

[9] IEC 61967-5: “Integrated circuits, Measurement of electromagnetic emissions, 150kHz to 1GHz - Part 5: Measurement of conducted emissions, Workbench faraday cage method”, 47A/615/CDV, 2001-05-18

[10] IEC 62132-5: “Integrated circuits, Measurement of electromagnetic immunity –Part 5: Measurement of conducted immunity - Workbench Faraday cage method”, 47A/542/NP, 2002-03-08

[11] IEC 61967-6: “IEC 61967-6: Integrated circuits, Measurement of Electromagnetic Emissions, 150kHz - 1GHz Part 6: Measurement of conducted emissions, Magnetic probe method”, 47A/608/CDV, Committee Draft For Vote, Date of circulation 2000-11-10

[12] IEC 62132-3: “Integrated circuits - Measurement of radiated immunity - Part 3: Bulk current injection method”, 47A/624/NP, 2001-12-07

[13] IEC 62132-4: “Integrated circuits - Measurement of radiated immunity - Part 4: Direct RF power injection method”, 47A/625/NP, 2001 -12-07

Robustness in analog design

Michel De Mey
AMI Semiconductor Belgium BVBA
Westerring 15
B-9700 Oudenaarde, Belgium
michel_demey@amis.com

Abstract

Integrated circuit designers are confronted more and more with the harsh environment of automotive systems. This "harsh" system environment is translated into electromagnetic fields. The susceptibility of the circuit to these fields and the emission of such fields by the circuit are specified. Electromagnetic compatibility is sometimes regarded as a skill, or an art, rather than a science, where empirical experience is mostly the basis for improvements. In this paper will be explained how EMC analysis can be introduced in a structured way into the design flow. Based on conductive precompliance test methods an analysis method is worked out that can be integrated in a normal design flow and can be used by the IC designers to verify "the robustness" of their analog design. As a demonstration of the analysis method some examples are worked out.

1. Introduction

The design teams dealing with the design of mixed signal integrated circuits for applications in e.g. the automotive environment, are confronted with translations of the "harsh" system requirements into design requirements for the integrated circuit.

These IC requirements are defined to guarantee electromagnetic compatibility of the IC into the system. A good definition of EMC can be found in the UK Defense Standard 59-41:

"Electromagnetic compatibility is the ability of electrical and electronic equipment's, subsystems and systems to share the electromagnetic spectrum and perform their desired functions without unacceptable degradation from or to the specified electromagnetic environment."

A. van Roermund et al. (eds.), Analog Circuit Design, 243-253.
© 2003 *Kluwer Academic Publishers. Printed in the Netherlands.*

One has on one hand electromagnetic susceptibility, EMS, which is the discussion on how sensitive a circuit is to the injection of noise. The degree of resistivity to such disturbances is a measure for the EMS quality of a system, and defines the ability to operate in an EM environment.

On the other hand one has electromagnetic emission, EME. EME of an electronic system describes the amount of electromagnetic disturbances that is generated by that system and that is sent into the environment. Every emission is a polluter of the EM environment. See also Figure 1

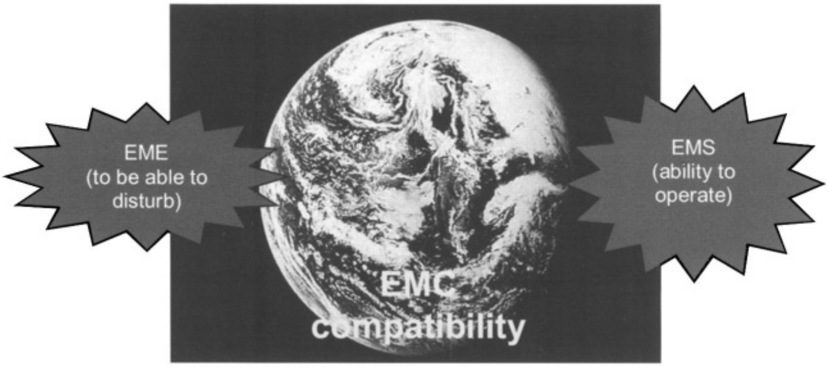


Figure 1

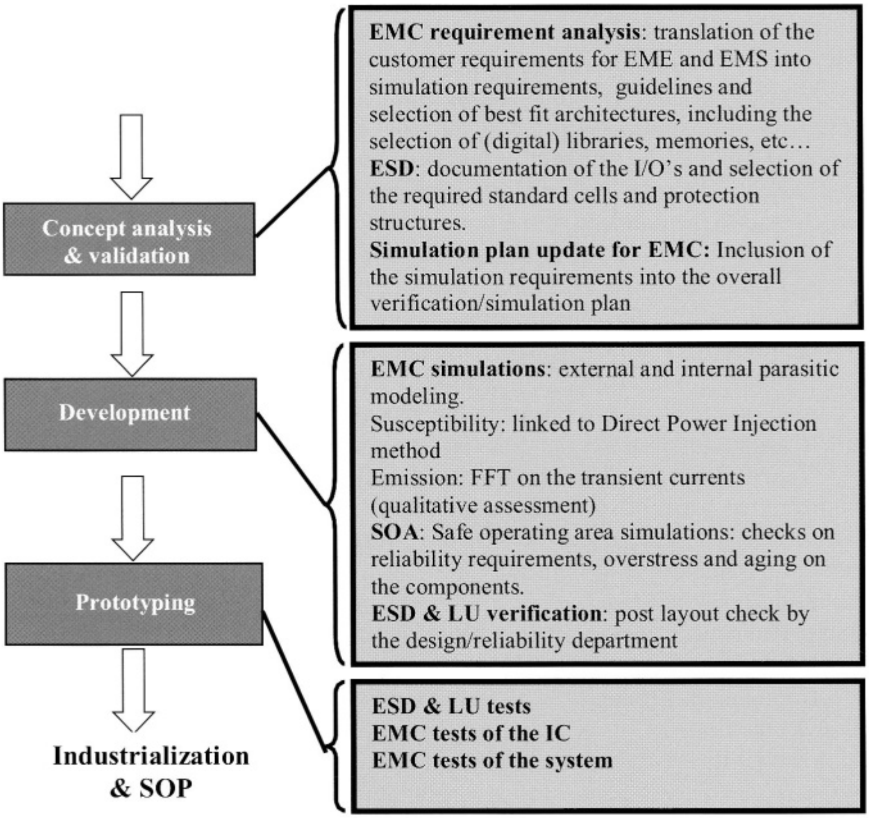
The EME and EMS requirements lead to “robust” design techniques. In this paper we will point out how the analysis of the “robustness” of the design can be introduced into the design flow, and how the different concepts and solutions can be analyzed versus their requirements.

EMC is a very broad field. The focus in this paper will be on the susceptibility analysis of analog design, rather than on the fields of emission reduction techniques in both analog and digital design and the ESD behavior.

2. Electromagnetic compatibility in the development flow

The analysis of the electromagnetic compliance and the definition of appropriate solutions needs to be handled as early as possible in the development flow of ASICs and ASSPs for the automotive market. If issues are detected late in the development (e.g. after prototyping) the degrees of freedom for finding a solution are strongly reduced due to time to market constraints. One can not completely change the IC nor the system anymore. Given the heavy cost pressure in these markets, also the cost of the BOM, the bill of material is crucial. The usage of filter-components like common mode chokes, surge clamps, etc... is limited to the bare minimum required. Therefore the analysis and selection of the appropriate design concepts in the system and the IC needs to be done as soon as possible in the development flow.

A development flow which is in line with the TS16949 automotive quality standard consists out of a “concept analysis and validation” phase, a “development” phase and a “prototyping” phase, prior to the industrialization and SOP, the start of production. In the Figure 2 the EMC related analysis is depicted.



EMC requirement analysis: translation of the customer requirements for EME and EMS into simulation requirements, guidelines and selection of best fit architectures, including the selection of (digital) libraries, memories, etc...
ESD: documentation of the I/O’s and selection of the required standard cells and protection structures.
Simulation plan update for EMC: Inclusion of the simulation requirements into the overall verification/simulation plan

EMC simulations: external and internal parasitic modeling.
Susceptibility: linked to Direct Power Injection method
Emission: FFT on the transient currents (qualitative assessment)
SOA: Safe operating area simulations: checks on reliability requirements, overstress and aging on the components.
ESD & LU verification: post layout check by the design/reliability department

ESD & LU tests
EMC tests of the IC
EMC tests of the system

Figure 2

During the “concept analysis and validation phase” the EMC requirements are studied and translated into a set of simulations to be performed on the circuit. During this phase an effective segmentation of the circuit is done. The signal path is analyzed and the requirements for common mode rejection ratio, power supply rejection ratio, ground rejection ratio, S/N ratio and bandwidth of the circuits are analyzed. The necessary EMC filters in the signal path as well as into the supply path are introduced. The required simulations to be performed and the assumptions on parasitic elements are included into the simulation plan. ESD requirements are analyzed and the necessary I/O cells and protections are selected. As will be demonstrated in the following section the I/O selection and the EMS requirements are tightly linked together.

During the development phase the necessary verification actions are executed, in line with the simulation plan. The design engineer looks for the best compromise between performance, cost and EMC requirements.

The prototypes are verified on system level and on IC level. The automotive industry, guided and driven by work groups like the German workgroup VDE-AK767-13-5, has translated the system level EMC tests into IC specific tests.

These tests are of conductive nature, which is a big advantage for circuit designers, since they work with currents and voltages rather than electric and magnetic fields. The conductive method deals with circuit elements like coupling capacitors, inductors, and transmission lines, rather than antennas. The assumptions which lead to the conductive method for EMS and EME are detailed below.

For EMS the following evaluation is made. The largest geometry found in an integrated circuit is the leadframe. The size of the leadframe is in the range of some centimeters or smaller. The dimensions of the structures on chip are at least one to two magnitudes smaller than the dimension of the leadframe. For the frequency range below 1 GHz this leadframe and a fortiori the structures on-chip are not regarded as efficient antennas for the reception of unwanted RF energy. It is the much larger cable harness and/or the traces of the printed circuit board which may be efficient antennas. Thus an IC gets the unwanted RF energy through those pins connected to wires of such cables. Therefore, the electromagnetic immunity of an IC can be characterized by conducted RF-disturbances (i.e. RF forward power) instead of field parameters as it is usually done in module and/or system testing. For EMS the requirements like described in the ISO 11452-2 and ISO 11452-3, dealing with shielded chamber measurements are translated into the conductive precompliance methods like the IEC 62132-4, Integrated Circuits - Measurement of Electromagnetic Immunity -150 kHz to 1 GHz, direct RF power injection method, the DPI method. [1]

For EME the following evaluation is made. The emission of an IC is generated by sufficiently fast changes of voltages and currents inside the IC. These changes drive RF currents inside and outside the IC. The RF currents cause conducted EME, which is mainly distributed via the IC pins by loops in the printed circuit board (PCB) and the cabling. These loops are regarded as the emitting loop antennas. In comparison to the dimension of these loops, the loops in the internal IC structure are considered to be small. See also [2]. For EME the requirements like the ones carried out according to the CISPR25, measured in the TEM cell are translated into the conductive methods, like the IEC 61967: Integrated Circuits, Measurement of Electromagnetic Emissions, 150 kHz to 1 GHz, part 4, measurement of conducted emission, 1 Ω / 150 Ω method. [3]

3. Analysis technique of the susceptibility of analog cells

Based on the direct power injection method a circuit analysis method for analog cells is defined. The direct power injection method couples by means of an RF amplifier and a capacitor Ccouple (typ. 6,8 nF) RF disturbances into the pins of the IC. The forward and reflected power are measured. The RF amplifier is terminated with 50 Ω. The method takes care that the external components to make the IC working are taken into account. The principle of a single pin power injection is depicted in Figure 3.

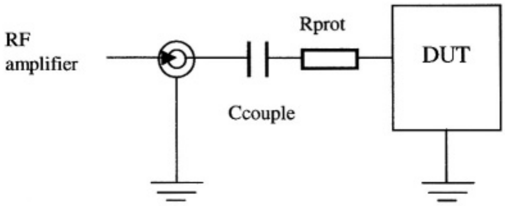


Figure 3

The default resistor value Rprot is 0 Ω. The value can be increased to 100 Ω if the functionality requires it. Note that also the supply pins are considered.

The set-up of Figure 3 is easily translated into a simulation schematic, which can be used to analyze the robustness of the circuit under test. See Figure 4

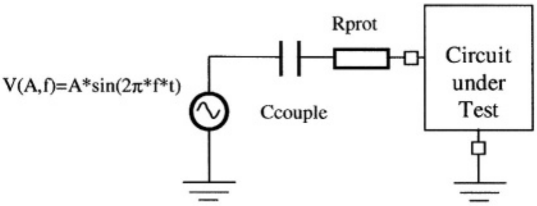


Figure 4

The circuit diagram and simulation stimuli represent the basic schematic. We will work out recommendations for the frequency f, the parameter A of the source and the accuracy requirements of the model of the circuit under test.

The frequency f exercised on the circuit under test should range from 1 MHz to 1 GHz to be in line with the DPI requirements. An appropriate selection of simulation frequencies is required in order to reduce the set of simulations to be performed in the time domain. When there are specific frequency characteristics of the external networks it is advised to select the frequencies with the lowest attenuation. When no specific information is available one can choose arbitrary:

f_1 = 1 MHz; f_2= 10 MHz; f_3= 100 MHz; (f_4= 1 GHz)

The frequency f_4 can in most cases be excluded based on the following evaluations. Robust analog design will strive for bandwidth limitation. The signal bandwidth in automotive signals is limited to 1MHz. Thus it is possible to include in the design of the I/O's a first or second order passive filter in order to reduce the bandwidth. Parasitic elements on the PCB and bondwires will also act as low pass filter elements at frequencies of the order of 1 GHz.

The amplitude A of the injected sine wave is defined by the EMS immunity level requirements, the input impedance of the circuit under test and the external circuitry of the pin to the "real world".

In case the input impedance $Z_{in}(f)$ of the circuit under test is bigger then 200 Ω, it can be considered as a high impedance versus the 50 Ω termination resistor of the RF power amplifier. Under these conditions a guideline for the amplitude A for simulations of automotive applications is given in the following Table 1. Note that the DPI target values are continuous wave values. (CW)

Level	**A (V)**	**DPI target (W)**	**Comment**
1	22	5	Direct connection of the I/O to the environment.
2	7	0,5	Direct connection of the I/O to the environment. Some external L-, R-, C-low pass filtering is available. Examples are signal conditioning input stages, sensor interfaces, …
3	2,2	0,05	No direct connection of the I/O to environment. Examples are interface chips, interfacing with sensor chips mounted on the same module, AD converters input stages, …

Table 1

In case the input impedance $Z_{in}(f)$ of the circuit under test is less then 200 Ω the amplitude of the voltage source will not be solely defined by the termination of the RF power amplifier. A big part of the RF power will be converted into current flowing into the pin of the circuit under test. The RF power will no longer be reflected by the circuit under test, but will be dissipated inside the circuit. Under these conditions it is advised to change the V-source of Figure 4 to an I-source, like depicted in Figure 5. Note that under these conditions the Rprot needs to be 0 Ω.

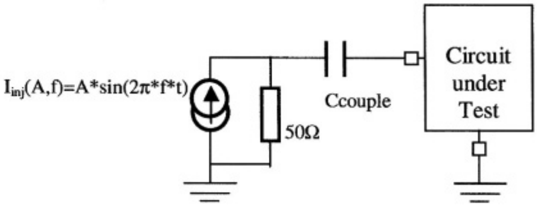


Figure 5

The amplitude of the current source I_{inj} is defined based on the following formula with P_{inj} the required immunity level, and Zin the input impedance of the circuit under test:

$$I_{inj} := \sqrt{\frac{P_{inj}}{\left(\frac{50\,\Omega\,Zin}{50\,\Omega + Zin}\right)}}$$

The quality of the simulations is determined by the quality of the models. This means also that parasitic elements need to be included into the simulation schematic. The simulation, being a transient simulation, requires accurate large signal AC models. Voltage dependency of junctions capacitors and MOS capacitors must be included in the model as well as all relevant parasitic transistors. As parasitic elements one should include all relevant junctions in the signal path and all parasitic capacitances and resistors. These elements can be extracted from the layout via dedicated tools. A point of attention is also to include all circuitry connected to the I/O into the simulation.

4. Interpretation of the simulations and circuit examples

The analysis of the simulations should focus on the effects of rectification and parasitic currents.

The drifts in the node voltages and currents versus the conditions in normal mode, also called rectifications, need to be analyzed. These drifts are caused by clamping or non linear, asymmetric behavior and can cause non linear behavior in the following stages due to saturation. They can even be interpreted by the next stage in the signal path as effective signal. The root cause of these rectifications lies in the asymmetries in the circuit. These result from mismatches in filter characteristics, asymmetries in the layout, or just by the large signal response of the circuit or devices. A typical example is the emitter or source follower circuit. See Figure 6

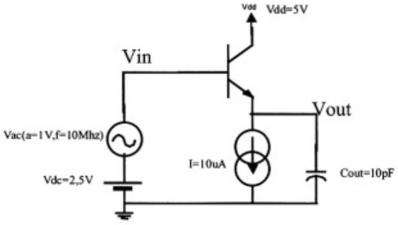


Figure 6

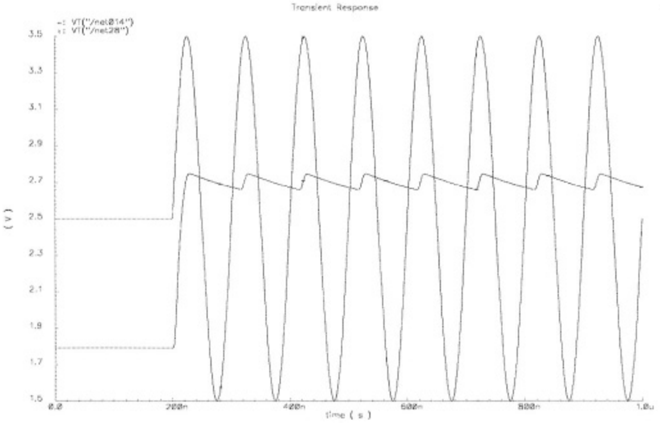


Figure 7

This widely used circuit has a large signal rectification effect, since the charging of the load is done by the active emitter or source follower, wile the discharge is done by a current source. This effect is demonstrated in the simulation result in Figure 7. The signal Vout is pumped up by the input from it's original value of 1,8V to a value 2,7V.

Another effect to monitor during the analysis are parasitic currents. Parasitic currents are characterized by current flow in parasitic devices. These parasitic elements, amongst which junction diodes and substrate PNPs, are not designed to conduct current, and are injecting current into the substrate. If not correctly captured by special layout techniques, these currents can be trapped in susceptible circuits, causing circuit malfunctioning.

One example of such structure, which causes clamping and rectification and injects parasitic current into the substrate is the well-known double diode ESD protection structure in the I/O. See Figure 8. The diodes of this structure clamp the injected signal and their associated parasitic transistors generate substrate currents. The parasitic devices are indicated on Figure 8 and can be understood by the cross sections in Figure 9.

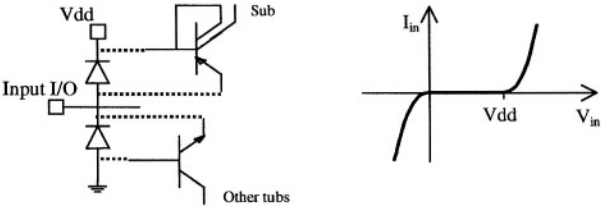

Vdd
Sub
Input I/O
Other tubs
I_{in}
Vdd
V_{in}

Figure 8

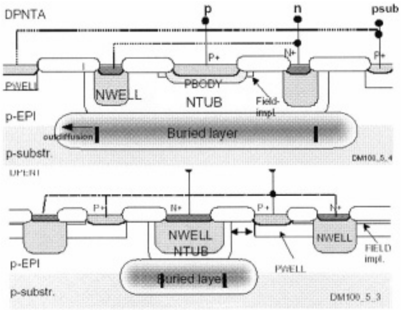

DPNTA
p
n
psub
P+
N+
P+
PWELL
NWELL
PBODY
NTUB
Field-impl
p-EPI
outdiffusion
Buried layer
p-substr.
DM100_5_4
P+
N+
P+
N+
NWELL
NTUB
NWELL
FIELD impl.
PWELL
p-EPI
Buried layer
p-substr.
DM100_5_3

Figure 9

An improvement for this structure is the I/O depicted in Figure 10 . It is made of two clamping structures triggering at higher positive and lower negative voltage in a symmetric way. V_{clamp} is typically 12 V. With this structure one can increase the immunity level and avoid rectification towards the supply rails. An improvement of the EMS behavior of an order of magnitude is demonstrated on silicon.

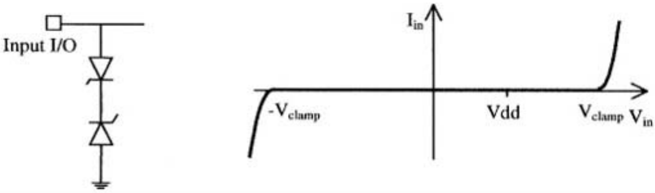

Input I/O
I_{in}
$-V_{clamp}$
Vdd
V_{clamp}
V_{in}

Figure 10

The analysis described can be used on top level circuits. As an example we have included here a simulation schematic of a high precision LED driver current regulation loop. See Figure 11. The schematic includes a dedicated model for the external components (see A), includes the direct power injection voltage source (see B). The layout parasitics are extracted in the resistor bank (see C), and observation filters (see D) are included to check the filtered voltages, which demonstrate the rectification (see E). Based on this simulation schematic the performance can be analyzed and where needed improved.

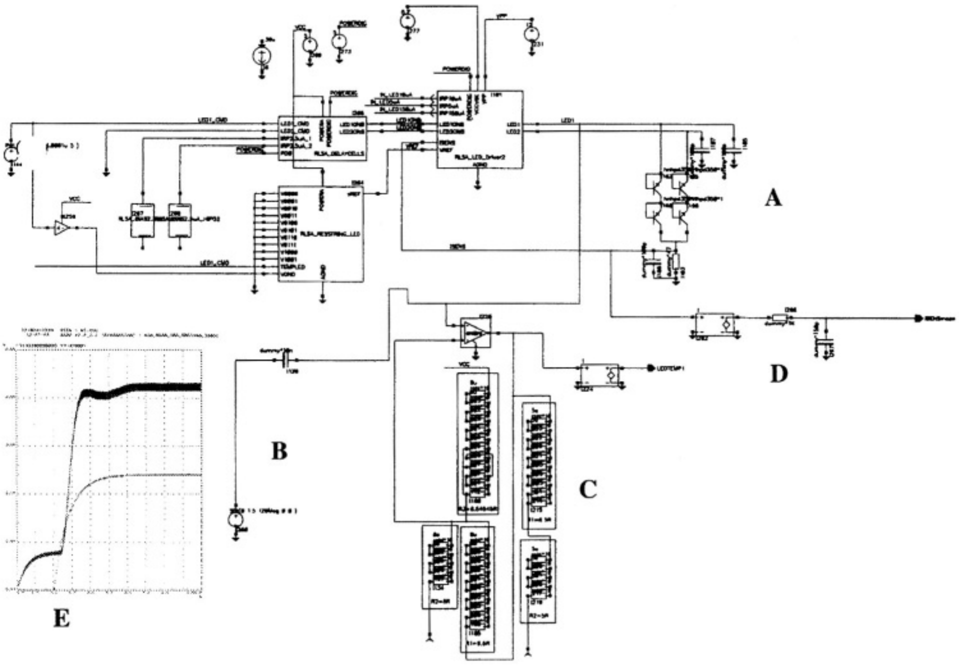


Figure 11

5. Conclusion

In this paper we have presented how electromagnetic compatibility is inserted into the design flow. The electromagnetic compatibility requirements, in the form of electromagnetic susceptibility and electromagnetic emission requirements, originate from the automotive system requirements and can be translated into conductive tests like described in the IEC 62132-4 and IEC 61967-4. These conductive tests have the advantage that they deal with voltages and currents, rather then E- and H-fields. Hence, these tests can be included by the automotive ASIC designers into their simulation test benches. An analysis technique for susceptibility of analog cells is worked out, and some examples are given.

6. Acknowledgements

The author wants to thank the colleagues of the product development organization of AMI Semiconductor Belgium BVBA and AMI Semiconductor Inc. for their contribution to this paper.

References

[1] International electrotechnical commission, IEC 62132-4, Integrated Circuits - Measurement of Electromagnetic Immunity -150 kHz to 1 GHz, direct RF power injection method, draft document in 2002

[2] M. Montrose, “EMC and the Printed Circuit Board: Design, Theory, and Layout Made Simple”, IEEE Press Series on Electronics Technology

[3] International electrotechnical commission, IEC 61967; Integrated Circuits, Measurement of Electromagnetic Emissions, 150 kHz to 1 GHz, part 4, measurement of conducted emission, 1 Ω / 150 Ω method.

This page intentionally left blank

MINIMIZING UNDESIRED COUPLING AND INTERACTION IN MIXED SIGNAL ICs.

Timothy J. Schmerbeck
JDS Uniphase, Data Communications Group, USA

Abstract

This paper surveys techniques and strategies aimed at minimizing coupling and interaction on mixed-signal IC's. Consideration is given to switching induced substrate coupling, power rail and i/o driver resonances, as well as near field capacitive and inductive interaction.

1 Introduction

Recent product demands for smaller size have been met using multi-chip modules, chip-on-board (COB), and deeper integration ICs. In some application areas, integration has approached a complete System On a Chip (SOC). SOC ICs are usually required to combine significant portions of analog circuits with the purely digital switching functions since the interface to most real world systems is analog. Current IC and packaging geometries have been miniaturized to the point that, at high speeds, even pure static CMOS logic designs are being limited by chip/package/card line to line crosstalk, output driver inductive di/dt switching noise problems, and power supply disturbances due to the RLC nature of the supply lines. Reference [2] shows that line to line coupling for rail to rail CMOS signals can exceed the static CMOS noise threshold of 1/2 Vdd for long wires. Adding analog or any circuitry with less noise margin than CMOS logic to these ICs is a very difficult task. This is also the motivation causing many to put sensitive analog and high switching rate digital on separate ICs. For example, in communications systems, it is still common to put preamplifiers on separate chips due to the low input signal amplitudes and their sensitivity to switching noise. For some noise sensitive classes of circuits, like these, this may be the only alternative; separate ICs mounted in a multi-chip module, separate packages, or COB packaging. For a sub-class of analog circuits, the analog can be made to coexist with the switching functions. Each year this analog sub-class is increasing and the number and type of logic circuits coexisting with them is expanding. This may be at the cost of strict partitioning of switching and non-switching functions, extensive special handling, special CAD tools, special semiconductor process, and a fully custom design effort. When the transmission line behavior of chip power busses is considered, noise can actually be minimized by selectively narrowing and/or widening connections from circuits to the power bus. Software driven power bus routing has shown itself superior to any manual efforts. Noise can be minimized by optimized, selective placement and sizing of on-chip decoupling capacitors. Again, a software driven solution to even decoupling capacitor placement versus manual distribution has shown itself

A. van Roermund et al. (eds.). Analog Circuit Design, 255-271.
© 2003 *Kluwer Academic Publishers. Printed in the Netherlands.*

superior. Availability of automated layout as well as signal and power routing CAD targeted specifically to analog and mixed signal designs has been limited since these tools are usually targeted to higher volume logic ASIC designs. Since a single-chip solution is often the smallest, lowest cost and lowest power implementation, the additional effort is generally justified. As the number of logic switching functions increases, the degree of difficulty merging the digital and analog functions will increase. Near field coupling between neighboring circuits, and coupling between widely separated circuits through the chip substrate and power rails are the big problems. As chip and package dimensions and clock frequencies increase, the wavelengths of the signals become comparable to chip and package interconnection lengths and this makes interconnections better antennas causing radiated emission problems from a single packaged IC. Radiated emissions problems can result from chip/package antenna lengths that are only 1/ 20th of the problem frequency wavelength or 1.5mm at 10GHz. Exceeding radiated emissions specifications usually occurs prior to manifestation of any actual functional concerns where the emissions perturb the design itself. Transmission line effects on silicon chips can't be neglected for 10ps rise times and wire lengths exceeding 600 microns. R, L, and C couplings become more significant with higher packing density. Also, the scaling up of substrate doping with constant field scaling laws makes physical partitioning of functions more difficult in simple bulk CMOS nwell processes. P+ wafer processes, with p- epitaxy, already see the effects. Triple well, mixed-signal CMOS, processes providing well isolated nmos devices in P substrates, are becoming more common to enhance circuit isolation capabilities. SOI processes are affected much less by bulk substrate doping and their built in oxide isolated wells greatly enhance circuit isolation capabilities. Deep trench isolated, SiGe Bicmos processes, achieve isolation capabilities equal to SOI processes.

As the total chip switching current increases it becomes very difficult to control inductive noise in the power lines. Card wires, package pins, bond wires, and IC interconnections all add inductance that cause fluctuation of the power lines and couple noise. The chip substrate acts as a collector, integrator and distributor of coupled noise on-chip. Very Large mixed-signal chips usually require low inductance packages, shielding/isolation structures, on-chip decoupling, custom on-chip substrate ties, n-well ties, power, and signal routing with physical partitioning of function, custom chip floor planning and every circuit noise rejection trick available [1]. Figure 1 Illustrates many of these problems for a chip wire bonded in a leaded surface mount package on a card.

Crosstalk and coupling through interconnects has been the subject of much research. Most integrated circuit designers today use RLC models of chip, package and board level interconnects to more accurately analyze their designs for crosstalk and supply bounce. In addition all chip interconnect R's, C's and sometimes L's and their coupling components are typically extracted by CAD software. However, the CAD extraction and subsequent simulation of noise coupling through the common substrate of mixed-signal systems has been neglected, mostly due to the difficulty in dealing with analysis of the substrate itself which is in effect a multidimensional interconnect connecting every transistor on the die with every other one. While mixed-signal designers have in the past employed certain crude models and device level CAD for substrate coupling in the verification of their systems, accurate parasitic extraction CAD to allow detailed simulation is now available from commercial CAD vendors such as Cadence/Simplex Substrate Storm or Cadence/Cadmos Seismic. The use of this CAD does significantly impact the resource required for a design.

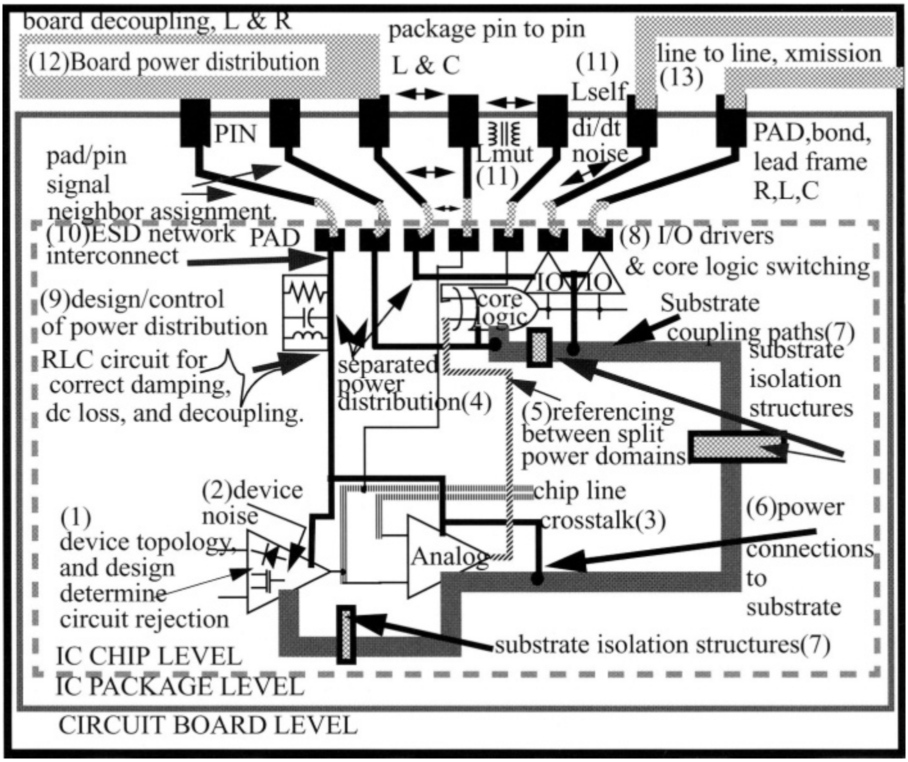

board decoupling, L & R
(12)Board power distribution
package pin to pin
L & C
(11)
Lself
line to line, xmission
(13)
PIN
Lmut
(11)
di/dt
noise
PAD,bond,
lead frame
R,L,C
pad/pin
signal
neighbor assignment
(10)ESD network
interconnect
PAD
(8) I/O drivers
& core logic switching
IO
IO
core
logic
(9)design/control
of power distribution
RLC circuit for
correct damping,
dc loss, and decoupling.
separated
power
distribution(4)
Substrate
coupling paths(7)
substrate
isolation
structures
(5)referencing
between split
power domains
chip line
crosstalk(3)
(2)device
noise
(1)
device topology,
and design
determine
circuit rejection
Analog
(6)power
connections
to
substrate
substrate isolation structures(7)
IC CHIP LEVEL
IC PACKAGE LEVEL
CIRCUIT BOARD LEVEL

Figure 1 Illustrative chip/package/card environment.

2 Sources of Noise and Methods of Coupling

Clearly the lowest level noise present on semiconductor chips is due to electronic device noises caused by the random movement of charges through resistances, across transistor junctions, and random fluctuations in the charge recombinations in surface states and the semiconductor bulk. The level of noise generated and coupled by thermal noise, avalanche noise, shot noise, and 1/f noise represents a minimum level in coupled noise and all other noise mechanisms treated are usually orders of magnitude worse than these without careful design. It is difficult enough to produce an amplifier with input referred noise less than 1nano-Volt/root Hz with device noise alone. Device noise control is accomplished mainly through optimum circuit design and topology with bandwidth limiting of signals. The circuit effects of chip thermal gradients, mechanical or piezoelectric stress, hot electron effects, and mobile ions such as sodium, can be considered very low frequency noise or noise coupling. Their control is usually accomplished with careful consideration of chip isotherms and mechanical stress lines, circuit design and biasing, balanced physical layout, and process control.

Clearly circuit voltage and current switching are the biggest noise sources on today's ICs and the simple, but often not practical, solution is to reduce the magnitude and frequency of current and voltage switching. Techniques that reduce switching power such as shutting down all switching functions or logic drivers when not in use also reduce coupled noise. There are also many clever techniques used to distribute switching events over longer periods of time or stagger switching events, and reduce the peak magnitude of current disturbances with reduced rise times. Many of these techniques do not reduce the overall switching energy but instead reduce the peak values of a disturbance. In general noise power will increase as the magnitude of switching events increase (signal swing, Vdd); The number of switching events increase (more gates, more transitions per clock cycle, more clocks); and the frequency of switching events increase (higher f). Many circuit performance specifications deal with the maximum deviation of a circuit from nominal performance versus a statistical average deviation. An example of this would be peak to peak timing jitter. This specification is more affected by peak levels of noise versus average of rms levels. For this reason, analog circuits and many digital circuits are most sensitive to the peak current or voltage noise disturbance values. Even though a single 50mWatt CMOS I/O driver has the same noise power as 50mWatts of internal core logic, the I/O driver concentrates the power on one switching edge and produces many times more peak noise on the power rails and substrate.

3 Contribution to Peak Switching Noise of Logic Family and Design Style

In general the logic family design styles with the least peak current or voltage noise disturbance values are those that employ small value, balanced, differential current steering or voltage switching. The worst are those that like CMOS are single ended, switch rail to rail, and concentrate very high currents at the voltage switching transition. Pre-charge logic like domino avoids the "shoot thru" current when both n and p mos switches are transiently on but adds many more clocks and has lower noise margin. Self-resetting logic adds the extra problem of variable frequency reset noise and many clocks.

Static CMOS, despite its wide-spread use, does represent near the worst choice as far as unbalanced, maximized peak current and voltage switching. Unfortunately, the most readily available logic libraries available today and the easiest to design in are static CMOS. This trend does not appear to be changing anytime soon. For chip designs containing large amounts of logic, it is usually preferable to try to partition and isolate the sensitive or low noise margin circuits from the higher noise margin, high switching rate, static and dynamic CMOS switching circuits on chip. This partitioning allows static CMOS to be used with little impact on sensitive circuits in another chip partition for lightly doped p- and Silicon On Insulator (SOI) as long as great care is taken in the design and partitioning. Highly doped p+ substrate processes are much more difficult to partition. This strict partitioning is not always possible so for very noise sensitive mixed signal designs, viable and practical alternatives to static CMOS do exist for logic on the same chip power rails as sensitive mixed signal circuits[4]. This is especially true at high frequency with high switching factor functional blocks that are the biggest noise producers and where the low static power of static CMOS is not as important. Note that logic coupled noise power to the chip substrate and chip power supply rails is proportional to the switching power of the circuits for a given chip/package model and a fixed logic circuit family. Lowering switching power lowers coupled noise power. For very slow switching rate circuits and near DC circuits static CMOS does not cause problems. At high switching rates the non-optimized or non-custom gate array and standard cell static CMOS circuits are very power wasteful and thus cause excess noise. Non-optimized or non-custom gate array and standard cell static CMOS circuits generally employ larger driver devices for global nets and result in much higher gate switching capacitance. Many modern static CMOS logic libraries are combatting this problem by providing a dozen or more gate sizes for each gate or latch type in the library. Software is then employed to minimize the size of each switching gate. Even fully optimized static CMOS circuits are power wasteful in comparison to some recent differential circuit families. Even when the static CMOS power is equivalent the “crowbar” or “shoot-thru” current developed when the NMOS and PMOS devices are on at the same time creates excess noise. Capacitive current coupled to the chip substrate or adjacent conductor or diffusion is proportional to the rise time of the voltage signal. Non-optimized or non-custom gate array and standard cell static CMOS circuits generally employ excessively larger driver devices for global nets and result in much higher rise times than are required.

4 Inductive Coupling

Table 1 shows calculated examples of self and mutual inductance of chip wires, wire bonds and 68 Pin PLCC lead frame wires together with examples of self inductive voltage bounce amplitudes and coupled mutual noise amplitudes. All but the last line in the table makes the pessimistic assumption that the only ground is at chip or card level so no pins are grounded. The inductance is reduced dramatically by the presence of grounded or ac low impedance pins/pads. The mutual inductive coupling and self inductive bounce voltage amplitudes become quite large even at moderate currents. This is why high switching and sensitive non switching functions can usually not share common or adjacent power or signal pins. Elimination of the small on chip coupling portion is possible with custom routed star power connections from switching and non-switching circuits to the common connection. In many analog circuits the on-chip mutual inductance coupling terms are not acceptable even for long chip wires. This coupling increases with each successive process technology generation due to thicker, narrower wires on a tighter pitch. These coupled amplitudes from inductive effects are easily exceeded by capacitive current coupling for higher impedance circuit nodes.

The mutual inductance, transformer coupled, voltages are most damaging on low impedance power supply nets that are unaffected by large capacitively coupled currents. For this reason, some on-chip decoupling is required to keep more of the supply transient current on-chip versus flowing through the bond and package inductances. The mutual inductance is used to advantage by placing CMOS ground and VDD pins adjacent. Because the currents are in opposite directions the mutual inductance subtracts from the self inductance.

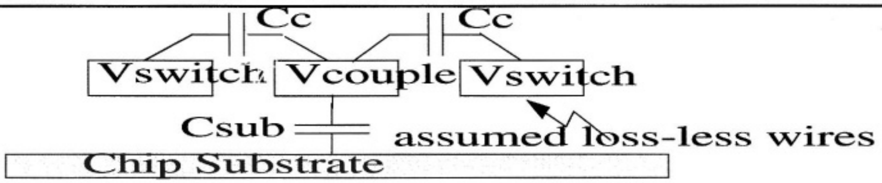
With package scaling from a 68 PLCC to a more modern 64 TQFP the area, lead pitch and package height all drop dramatically (5.85, 2.26, 2.7 times respectively). The resulting reduction in self C, mutual C, self L, mutual L for a corner lead drop 3.14, 3.44, 1.67, and 1.55 times respectively. However note that the mutual C and L as a percentage of the self C and L remain basically unchanged. The C-mutual/C ratio drops from 40% to 36.4% and the L-mutual/L increases from 53.8% to 57.8%. The percentage mutual coupling terms will not be reduced significantly by surface mount package scaling. Also frequency increases are reducing the benefit of self inductance drops in more advanced packages. In the limit of COB the coupling is that of a single wirebond (~1nH self & 45% mutual L) or a single flip-chip bump (~50pH self & very small mutual L.)

Conditions:	self-L(nH)	mutual L (nH) corner lead	mutual L coupling due to 10mA RMS @ 50MHz	mutual L coupling due to 100mA rms@ 50MHz	self-L bounce due to 100mA rms@ 50MHz (in mv)
1	0.42	.037	0.16 mV	1.6 mV	18.6
2	3.0/2.24	1.4	6.2 mV	62 mV	133/99.5
3	11/ 7.9	4.2	18.7 mV	187 mV	489/351
4	14/ 5.0	5.6	24.9 mV	249 mV	622/222

Table 1 *Inductive coupling from chip wires, bond wires, & 68 pin PLCC package leads. Condition 1: 1mm long chip wire with 3um width & space at 1um above the silicon surface. Condition 2: Wirebond wire only with (corner/centre bond). Condition 3: Lead frame wire only with (corner/centre lead). Condition 4: Total corner lead with (no nearby grounded pin)/(every other pin grounded).*

5 Capacitive Coupling

Figure 2 shows the capacitive coupling environment for wires immediately above the silicon substrate versus wires at higher levels. For lower metal levels there is less line to line coupling but more coupling to and from the substrate. As wire pitches and widths both shrink, while wire thicknesses increase, line to line coupling levels increase. For large chips with an inverting logic family like CMOS, there is, on the average, close to one rising signal for every falling signal. If the output loads to ground and Vdd are balanced, the total chip average of injected current into the substrate is small with the net effect that the substrate RLC circuit is stimulated. It is also possible to get locally higher injection due to poor local cancellation. For most static CMOS designs, interconnect and gate capacitive loads to ground are larger than to Vdd by a factor of as much as 4. Because of this, direct output coupled noise injected into the substrate can possibly exceed RLC resonance noise depending on the balance between load to ground vs supply. This is usually the case for CMOS chip I/O drivers. On smaller chips, a large area, single ended, locally interconnected, clock net routed on the first interconnect level above the substrate, can possibly inject large locally un-cancelled "output coupled" noise into the substrate. For this reason it is better to route on higher metal levels or use closely routed balanced out of phase clocks to reduce or cancel noise coupled to substrate. Often an isolated n well tied to a non-sensitive supply line is routed under a noisy clock line to prevent coupling to the chip substrate in a sensitive analog circuit area. Also, separation or shielding is necessary between noisy and sensitive wires. It is also necessary between the chip substrate or large p or n wells and noisy or sensitive wires. Since the separation between interconnect wire levels is approaching the separation distance between wires on the same interconnect plane, both vertical and horizontal shielding is required. Since wires on adjacent vertical wiring planes are usually automatically routed at 90 degree preferred directions it is only custom or "wrong way" wires that are of key concern. An interconnect wire, poly silicon or diffusion tied to AC ground potential offers the maximum noise rejection for a given chip area if the shield can remain uncontaminated by the noise it is shielding from. If this can not be guaranteed, a physical separation of noisy and sensitive net is the best rejection solution. In a 1 um CMOS process, leaving one empty wiring channel between two coupling wires provides 20 dB of noise attenuation. In a 0.18um process it will take roughly 3 empty wiring channels to achieve the same attenuation of 20dB.



Both Vswitch signals are in phase. First metal over diffusion.

Process Photolithography	Cc	Csub	Coupled % of switching voltage swing
1 micron	0.03 pF/mm	0.17 pF/mm	26% of Vswitched
0.5 micron	0.05 pF/mm	0.07 pF/mm	59% of Vswitched
0.18 micron	0.09 pF/mm	0.03 pF/mm	85% of Vswitched

$$\left((\Delta V_{\mathrm{coupled}}) \propto \frac{(\Delta V_{\mathrm{switched}})2C_c}{C_{\mathrm{sub}} + 2C_c} \right)$$

This equation relates the limit of line to line capacitive coupling and depends on process geometries.[2] Inductive coupling will be 50 to 100 times smaller for the same chip wire geometry and the same characteristic impedances. The equation above gives the coupling limit for long wires regardless of the line driving or coupled source impedance or the signal rise times.

Figure 2 Capacitive Coupling Scaled with process photo

6 Effect of Power Bus Structure on Noise coupling

The power bus that interconnects the switching and non-switching functions on the chip is a major source of coupling between widely separated circuits. Switching return currents take the path of least impedance which is often an on-chip path through the power rails. The presence of on-chip decoupling capacitance distributed on the bus as well as resistive damping can lessen the amount of bus fluctuation due to these switching currents. Figure 3 shows a single rail grid power connection scheme where the switching and non-switching functions have been strictly partitioned to use separate chip power pads. This is usually required due to the large inductive bounce voltages determined previously. It is not always possible to strictly partition switching from non-switching functions and in this case the STAR power routing gives less coupling at the expense of a totally custom power bus. If some on-chip power supply decoupling capacitance can be provided the split-grid power bus scheme can give superior performance without a custom bus design.

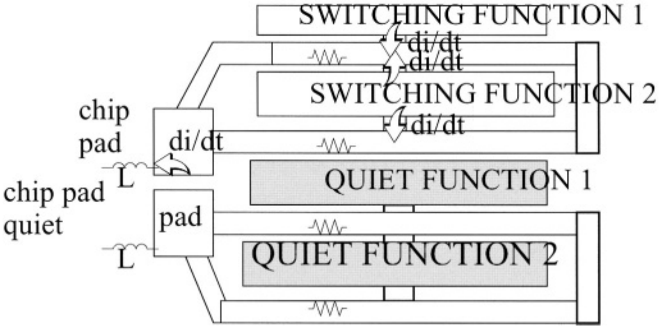


Figure 3 STAR power feed using Grids. (Best of Star and Grid)

Reference [7] describes a fully automated power bus routing software system that routes the bus to achieve minimum bus resistance and coupling while minimizing chip area used by the bus. A performance function needs to be specified and the coupling mechanisms need to be quantified for the particular chip process. Clearly this is the approach needed to achieve an optimized power bus routing. One of the problems that arises with a separate power bus and pads for switching and non-switching functions is communicating signals between the multiple power domains. Because the separate power buses have different noise waveforms superimposed on them, there is a time varying difference between the bus voltages in different power domains. Circuits that drive voltages across power domains as well as circuits that receive voltages from another power domain must contend with this time varying bus difference voltage. Figure 4 depicts the situation where a CMOS signal driven from the quiet analog power domain, on the left, must be received by an inverter that resides in a noisy switching power domain. The finite rise time of the analog inverter output causes jitter in the received signal since the receiving inverter switching threshold is changing with time, relative to the received signal, due to its supply bounce. This is a common situation since often a clock is generated by an analog phase lock loop circuit and communicated to the switching power bus. The jitter can be reduced by decreasing the rise/fall times of the analog output signal or reducing the voltage bounce on the switching supply. This problem often requires that clock generation circuitry reside in the same power domain as the circuitry that is driven by the clock. The problem can also be solved by passing differential signals across the power domain. This complicates the driver and receiver design since CMOS circuits are typically single ended in topology. Special differential receiver circuits are usually employed followed by a differential to single ended converter.

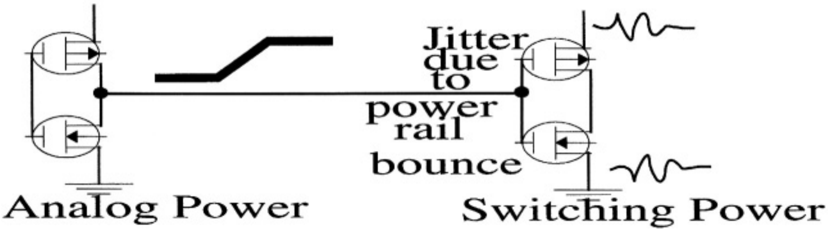


Figure 4 Communication between separate buses. (quiet to switching).

When a signal is communicated from the switching power domain to the non-switching or analog power domain, the same jitter situation occurs and needs to be fixed in the same way: In this situation a differential driver or two out of phase single ended drivers can partially cancel the noise current coupled to the analog power rails when the digital driver switches. Note that an unbuffered logic control signal to an analog switch carries the switching power supply bounce and resonance even when it is static or between switching events. The noise signal couples through the switch capacitance and enters the analog circuit. The signal coupling can be reduced by including one or more small series inverters in the analog power domain to remove the switching supply bounce signal. The problem with this approach is that if there are many signals that need to be buffered in this way, there will be large amounts of switching current added to the quiet bus when these buffers switch. However, it is still usually preferable to couple noise into the supply rails instead of a sensitive circuit node. If the signal entering the analog power domain from a switching domain is a slow speed control signal, it is often possible to insert a series resistor. This added series resistance will further reduce the amount of noise current coupled to the analog domain power rails. This added series resistance will also tend to isolate the two power domains. Note in Figure 4 that the capacitances of the inverters and the channel resistances of the driver and receiver devices effectively connect the two power domains with a series RC circuit. The more signals that cross the two power domains the lower the R and the higher the C of the linkage. If a large number of wires cross the two domains the two power domains are effectively the same. This current injection to the supply rails, is still preferable to currents injected directly into the p- substrate.

7 RLC CARD, PACKAGE AND CHIP RESONANCE

The power rails of an IC generally have the most capacitance to chip substrate and to each other due to the large number of devices connecting to them. This large capacitance together with the inductance and resistance from the chip, bonds, and package leads form multiple RLC resonant circuits with their own distinct natural frequencies. When stimulated by on-chip voltage and current transients, these RLC circuits will add their own characteristic noise frequencies. The sharpness of resonance and the resonant amplitude will depend on the Q of the system. The Q will vary with CMOS switching activity since the number of "on" gates or inverters partially determines the resistance or damping between the supplies. Tuning is usually required to prevent a power bus resonance from lining up with a system clock frequency. Tuning of the resonant frequencies is possible by controlling the amount of on chip capacitance and off chip inductance. Damping the RLC transient response can be accomplished by reducing the time constant 2L/R of the RLC circuit. This can be done by increasing the bus DC resistance or decreasing the system inductance. A variety of ways exist to increase on-chip capacitance but it is not easily reduced. The resonant frequency is usually raised and the damping increased by lowering the packaging inductance. This can be done with better packaging or increasing the number of package leads assigned to the power rails.

8 Off-chip driver RLC Coupling

The IC off-chip drivers also form an RLC resonant circuit which can be dominated by both off-chip capacitance and package inductance. This resonance is generally at a much higher frequency than power supply resonances and so coupling is greater through its capacitive connections to the substrate. The off-chip driver output voltage transient will ring with the characteristic frequency

$$f = \frac{1}{2\pi\sqrt{LC}}$$

where there can be a different frequency for the Vdd return loop from the ground return loop. If a large on-chip decoupling capacitance exists from Vdd to ground the frequencies will be the same since the capacitance puts the loops in parallel. Because the capacitances involved in the output loop arc the driver and load capacitances which are usually much smaller than in the supply loops the resonant frequency will be much higher. This resonant voltage ring on the output couples through the circuit capacitances to add to the noise integrated on the supply rails and substrate. The resonant voltage ring or any reflection transients on the output will also couple capacitively through ESD protect diodes and directly if ringing above the Vdd or below the ground rails transiently occurs. This ring is heavily dependent on the load and the rise and fall lime of the driver. Controlled rise time output drivers help to control the typically 4 to 1 process and environment variation in standard CMOS driver rise and fall times. On low performance drivers a series resistance can be inserted to damp the ring. High Performance driver design requirements usually don't allow much tuning of the I/O driver output RLC circuit but placing large drivers close to their power return leads limits the inductive loop area for reduced radiated emissions as well as best performance. The use of balanced current steering drivers or those with controlled rise time and reduced voltage swing together with staggering switching of drivers on a bus limit or partially cancel the di/dt and dv/dt to the substrate and chip power rails for lower coupled noise amplitudes.

9 Substrate Coupling

The chip substrate acts as a collector and distributor of noise on the IC. The amount of coupling varies depending on the structure and doping of the chip substrate as well as how it is tied to its assigned voltage potential on chip. Essentially every chip voltage transient on chip signal wires, I/O pads, and power rails is capacitively coupled to the chip substrate. This includes energy from card reflections back to I/O pads and transmitted to the substrate via I/O protect devices. Experiments have determined that the noise energy coupled to and from the substrate is proportional to the total chip switching power as well as to the logic power rail inductance and chip substrate tie inductance[1]. Coupled noise peak voltages are frequency dependent, unlike the coupled energy, due to constructive and destructive interference of the various frequencies. Noise peak voltages increase as a fixed number of logic transitions are concentrated over a narrower time interval even though the overall switching power is fixed and the total coupled noise energy remains unchanged. The switching of a capacitively loaded I/O driver represents an example of concentrating switching power in a narrow time interval. This results in much larger coupled peak noise voltages per milli-watt of switching power for an off-chip I/O driver versus a group of smaller drivers not exactly aligned in phase. In as much that power supply voltage, temperature, and frequency change switching power they also change substrate coupled noise energy.

10 Coupling in Wafers with p+ Bulk and p- Epitaxial Layer

Highly doped p+ bulk wafers with grown epitaxial layers are the second most common wafer type after p- bulk and the easiest to analyze, but usually offer the worst analog to digital coupling environment. An exception to this exists, at moderate to low frequencies, if the p+ bulk is very highly doped (~.006 Ohm-cm) and can be used as a pseudo ground plane under the epitaxial layer. This usually involves a low inductance and resistance connection to the entire chip backside via a metallization. The wafer must be backside ground to thin the bulk material to be less than the skin depth for the entire problem frequency range. At 100 MHz a 0.006 Ohm-cm, 8 inch, 750um thick un-thinned wafer will be about two skin depths thick but can be thinned to about 200um. The low inductance connection to the card can be achieved with chip-on-board packaging which uses conductive epoxy or solder to electrically connect the chip backside directly to the card ground or one of several surface mount leaded and lead less packages with exposed die-pad bottom side metal. The backside contact can also be used for the ground power bus eliminating the need for topside ground bus chip pads or bonds. If a backside contact is not used the substrate may be lied via topside bond wires to ground potential. Because of the spreading resistance due to the p channel stop in the p- epi, a single 1 mm long and few micron wide contact can make a connection of only a few Ohms to the underlying p+ bulk. This enables a few chip peripheral contacts not connected to any chip ground to be bonded out to eliminate the need for other substrate ties on chip ground. This prevents the direct resistive connection of chip power rails to chip substrate with the remaining substrate coupling mechanism being capacitive. There is no latch-up risk with only a few large substrate ties since the p+ bulk forms a low resistance connection to any place under the p- epi. If extra bond pads are not available to tie the substrate the next best alternative is to tie the substrate to a quiet ground bus. Note that substrate contacts tied to the ground bus are needed to make effective use of the natural decoupling capacitance provided by nwell to substrate junctions of CMOS processes. For this reason substrate contacts are usually only removed from analog ground busses where added decoupling capacitance can be deliberately added to make up for lost nwell capacitance[6]. Tying the analog substrate to switching ground is usually the worst alternative. Experiments with p+ bulk/p- epi wafers show that the separation of noise transmitter from noise receiver makes almost no difference in coupled noise due to the low resistance bulk substrate. The substrate's low resistance makes most current flow through the epi region vertically and renders n well guard rings basically ineffective from shielding a device from other than its nearest neighbor. p+ guard rings, when spaced close to the epi thickness away and tied to a non contaminated ground, can reduce substrate coupled noise to an NFET as much as 30% [3]. P devices in an nwell experience the least coupling from the substrate. Guard rings also have the ability to equalize coupling to matched devices. This is often desirable when noise can't be completely eliminated but can be rejected if it is common mode. Figure 5 illustrates this for both P+ and bulk P- wafers.

Reference [5] describes an analysis of an operational amplifier input composed of two large PFET devices in an n well of a bulk P- CMOS process, but the concepts apply to both P- and P+ substrate types. By using the LAYIN tool (now in Cadence/Simplex Substrate Storm CAD), described in reference [5], it is possible to plot the coupled noise signal path from noise source to the PFET input devices. The starting layout showed that the noise coupled differently to each of the two input PFET devices since one device was closer to the noise source and because the surrounding layout of devices was not exactly symmetrical. A P+ guard ring added to surround both input devices served to reduce the asymmetry in coupling and reduce the level of coupled noise. However, there was still an asymmetry in the coupling to each device. An n+ nwell contact completely surrounding the outside of the n well containing the PFET input devices was necessary to distribute the coupling to the n well and make the coupling almost completely common mode to both devices. A second p+ guard ring, outside the n well, would also reduce and further equalize the coupled noise. All of these guard ring structures increase the circuit layout area required and therefore increase the required chip area. In addition to equalizing coupling to matched devices, guard rings help to minimize body effect and capacitive coupling. The p+ guard rings help to equalize the potential of the substrate around an NFET device and reduce body effects as well as bulk to source coupling if the NFET source is referenced to the same ground that is tied to the p+ substrate guard ring. Likewise, the n+ ring inside an n well helps to equalize the potential of the n well. Processes that have buried p+ layers in the substrate and buried n+ layers in the n well to help latch-up also help equalize the potential (minimize the potential drop) across the substrate and n well in a local region beneath a device. To first order the coupling from the substrate to a circuit node in P+ wafers is proportional to the capacitance to substrate from that node and the impedance of the coupled node:

$$V_{coupled} \propto V_{substrate} \times C_{(node)-(substrate)} \times Z_{node}$$

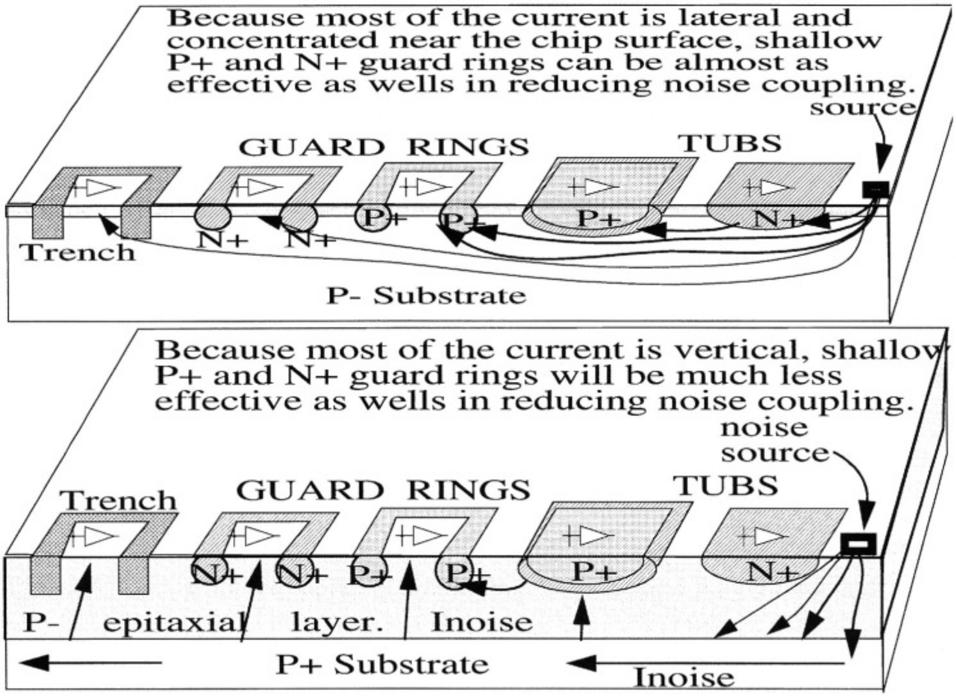

Because most of the current is lateral and concentrated near the chip surface, shallow P+ and N+ guard rings can be almost as effective as wells in reducing noise coupling.
source
GUARD RINGS
TUBS
Trench
N+
N+
P+
P+
P+
N+
P- Substrate
Because most of the current is vertical, shallow P+ and N+ guard rings will be much less effective as wells in reducing noise coupling.
noise source
Trench
GUARD RINGS
TUBS
N+
N+
P+
P+
P+
N+
P- epitaxial layer.
Inoise
P+ Substrate
Inoise

Figure 5 Guard rings and well isolation in P- Bulk & P+ substrates.

As a result, substrate noise coupled to PFETs in an n well tied to low inductance power will be lower than for an NFET in the epi. Triple well processes provide an isolated nfet from the substrate. Because of the large n well area with the substrate and the large capacitance between the power rails it will be difficult to get a “quiet” tie point for a large number of n wells. Likewise, a poly silicon resistor will have less noise coupled from substrate than a diffused resistor in the epitaxial layer. The most effective noise control techniques with p+ bulk with p- epitaxy processes is lowering switching peak power, minimizing power and substrate inductance, adding on-chip decoupling capacitance, avoiding substrate ties on both analog and digital ground busses, tuning power resonant frequencies away from frequencies of interest, well isolating sensitive circuits, and physically partitioning the power and associated package pins among quiet and switching functions.

11 Noise Coupling Control in Bulk p- Wafers

Bulk p- wafer CMOS and BiCMOS processes are the most common in use today. Typical resistivity is 10 Ohm-cm and ties to substrate are required to be distributed across the chip to prevent latch-up. Some processes use buried p+ layers to relax this requirement but the treatment is essentially the same. Backside contacts are not sufficient to prevent against latch-up alone even when the chip is thinned to 200 microns thick. Because of the area penalty to route a separate substrate tie bus and the desire to keep NFET source and body/bulk at the same potential to reduce noise in MOS devices, substrate contacts are usually tied to each ground bus on the chip. A large wide n well is commonly placed around partitioned non-switching power areas of the chip. The n well breaks the p surface channel stop implant region to provide a higher resistance path between the switching and non-switching power domains on chip. Due to the nwell capacitances the isolation resistance has a 3db roll off frequency of only a few GHz. If the p channel implant is blocked with an added block or moat mask, which is becoming more common, the native substrate resistance has a -3db roll off frequency of over 10GHz[5]. Note that when the chip substrate is split, in this way, much of the return current must flow off-chip. Often the currents are routed partially through the package die-paddle by bonding both analog and digital ground lead frame pins to the die-attache area under the chip. This reduces radiated emissions by shrinking the return current loop area. Measurements and simulations show that in a 10 Ohm-cm bulk substrate the current flow in the substrate is much more uniform with all significant currents flowing within about a 200 micron depth from the silicon surface in chips less than 6 mm on a side. The p channel stop at the surface has doping levels of typically three orders of magnitude higher than the bulk substrate so much current flow is at the surface. Because of this shallow lateral current flow near the surface, shallow n well guard rings are effective at diverting current flow downward into the bulk material and away from protected devices. Likewise shallow surface p+ guard rings can better control local substrate potential around devices due to the non vertical current flow and higher resistance of the bulk substrate. Shielding from shallow surface p+ contact guard rings is dramatically improved by almost an order of magnitude relative to the p+ bulk/p- epi substrates [3]. This is depicted in Figure 5. The coupled noise between resistively separated switching and non switching power domains decreases almost linearly with the separation resistance. However, because each domain shorts the substrate within, via ties distributed on its ground bus, not much benefit is obtained by moving a circuit from the edge to the centre of a power domain; much the same as with p+bulk/p- epi substrates. The improvement of devices in a well over devices in the bulk substrate (PFET vs standard NFET) is not as pronounced due to the lateral current flow. When multiple separated power domains are used it is very difficult to predict how the noise from a switching domain will distribute to the other domains without detailed simulations of the chip geometries.

12 Substrate Attenuation Structures

Figure 6 shows a technique that combines separation of switching from non-switching power domains and a method to attenuate coupling between the two domains. The situation is depicted for a bulk p- substrate. The separation or moat region can be built using an n well or substrate block/moat mask. In the moat separation region, p+ diffusions connected to metal strips and separate bond wires arc used to provide attenuation in the isolation or moat area. Figure 6 also shows the simplified electrical equivalent circuit. Zan is the equivalent impedance of the non-switching or quiet bus connections to substrate with the equivalent chip/package capacitance and inductance. Zn is the equivalent impedance of the switching or noisy bus connections to substrate with the equivalent chip/package capacitance and inductance. Rs1, Rs2, and Rs3 are the near-surface substrate resistances of the three separate moat isolation regions. Rc1 and Rc2 are the contact and metal resistances of the p+ regions that break the moat isolation area into three pieces. Likewise L1 and L2 are the inductances associated with the two contacts from bond wires and package. Rsd1,2,3 are the deep shunting resistances from node A to node D due to the current flowing deeper in the substrate than the p+ contact diffusions. Rcd1 ,2 account for the poorer effectiveness of the contacts at points B and C in collecting the deeper substrate currents.The circuit clearly provides attenuation of the noise at point A before reaching point D with low enough values for L1, L2, Rc1, Rc2 and high enough values for Rs1, Rs2, Rs3. If Rs1,2,3, are large and Rc1,2 and L1,2 are small the attenuation is large. Un-shown shunt capacitance across Rs1,2,3 inherent in the isolation moats and due to crossing signal wires limits the usable frequency range of this technique

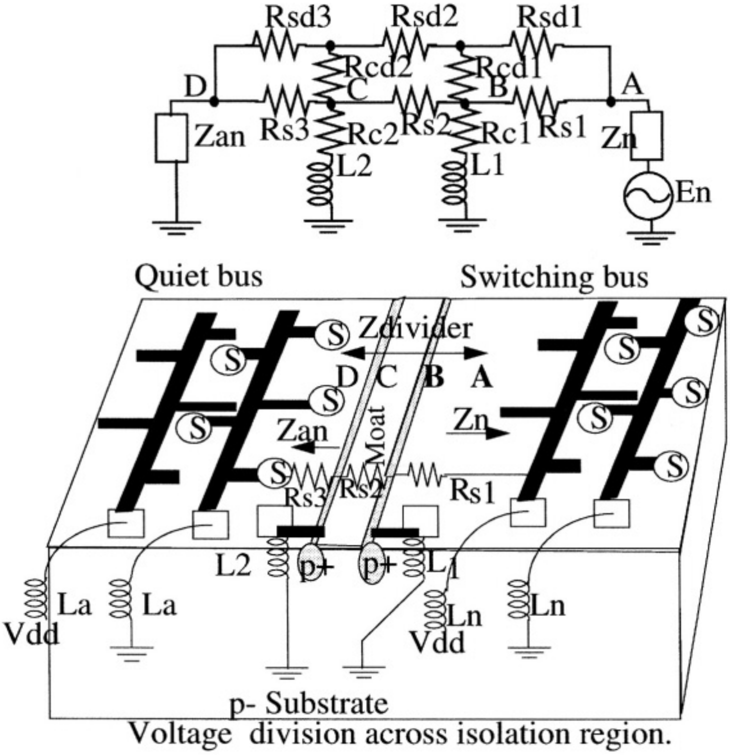

Rsd3
Rsd2
Rsd1
Rcd2
Rcd1
D
C
B
A
Zan
Rs3
Rc2
Rs2
Rc1
Rs1
Zn
L2
L1
En
Quiet bus
Switching bus
S
Zdivider
D
C
B
A
Zan
Moat
Zn
Rs3
Rs2
Rs1
L2
p+
p+
L1
La
La
Vdd
Ln
Ln
Vdd
p- Substrate
Voltage division across isolation region.

Figure 6 Substrate Splitting (voltage divider across substrate).

Figure 7 shows how to obtain very high isolation using multiple wells. In this example quiet and switching functions are placed in their own n well. Each n well also has dual well capability within itself making the overall process triple welled. Trench isolation is then used between wells to guarantee low capacitance between wells.

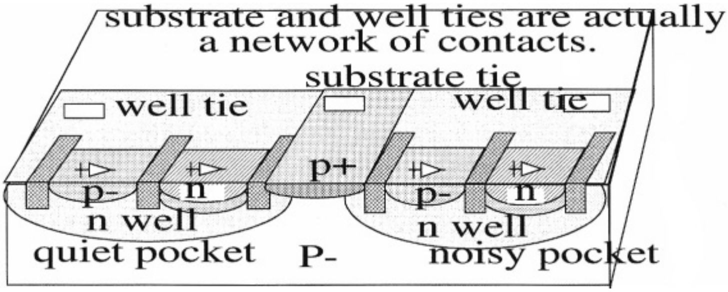

substrate and well ties are actually
a network of contacts.
substrate tie
well tie
well tie
p+
p-
n
p-
n
n well
n well
quiet pocket
P-
noisy pocket

Figure 7 Multiple substrate well isolation.

Figure 8 shows isolation capability available with SOI or Silicon On Insulator technology. SOI, Silicon on insulator processes, have the potential for almost separate chip levels of channel isolation between functions on chip. The switching and non-switching functions can be partitioned in diclcclrically isolated islands on the semi-insulating substrates. This technique can deliver the best isolation possible on a common chip substrate. Note that each island must be tied to a potential to collect the capacitive currents injected across the dielectric trench. SIMOX SOI wafers where oxygen is implanted below the silicon surface to produce a layer or oxide under a thin uniform silicon layer are achieving commercial success today. Most of the noise problems discussed will be reduced by orders of magnitude for SOI wafers.

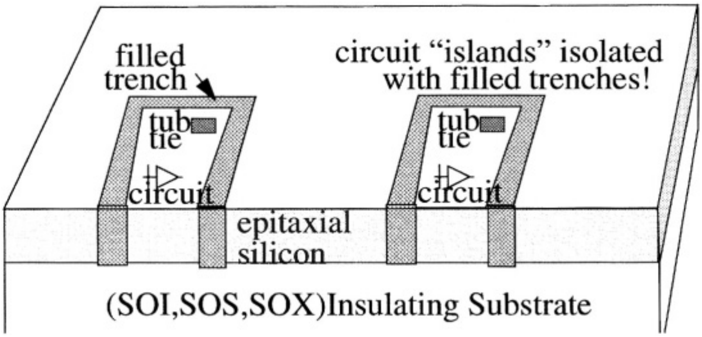

filled
trench
circuit "islands" isolated
with filled trenches!
tub
tie
circuit
tub
tie
circuit
epitaxial
silicon
(SOI,SOS,SOX)Insulating Substrate

Figure 8 Ideal situation is an insulating substrate.

13 Conclusions

This paper has surveyed practical aspects of key analog and analog/digital interaction problems.

14 References

[1] Nishath K. Verghese, Timothy J. Schmerbeck, David J. Allstot, Simulation Techniques and Solutions for Mixed-Signal Coupling in Integrated Circuits, Kluwer Academic Publishers, Co., Norwell, MA, 1994.

[2] Laszlo Gal, "On-Chip Cross Talk - the New Signal Integrity Challenge," In Proceedings IEEE Custom Integrated Circuits Conference, pp. 12.1.1-4, San Diego, 1995.

[3] D.K. Su, M.J. Loinaz, S. Masui and B.A. Wooley, "Experimental Results and Modeling Techniques for Substrate Noise in Mixed-Signal Integrated Circuits," IEEE Journal of Solid State Circuits, vol. 28, no. 4, April 1993.

[4] D. J. Allstot, "Low Noise Digital Logic Techniques", In Proceedings IEEE ASIC Conference, pp. T13.1.1 & T13.1.2, September 1991.

[5] Francois Clement, Computer Aided Analysis of Parasitic Substrate Coupling in Mixed Digital-Analog Integrated Circuits, PHD Doctorate thesis, EPFL, Lausanne, Switzerland, October 1995.

[6] Matt Felder and Jeff Ganger. Analysis of Ground-Bounce Induced Substrate Noise Coupling in a Low Resistive Bulk Epitaxial Process: Design Strategies to minimize Noise Effects on a Mixed-Signal Chip," *IEEE Transactions on Circuits and Systems-II,* vol. 46, no. 11, Nov. 1999, pp. 1427-1436.

[7] B. R. Stanisic, R. A. Rutenbar, and L. R. Carley, "Power Distribution Synthesis for Analog and Mixed-Signal ASICs in RAIL," In Proceedings IEEE Custom Integrated Circuits Conference, pp. 17.4.1 - 17.4.5, May 1993.

This page intentionally left blank

PART III: LINE AND BUS DRIVERS

In this chapter on wired line and bus drivers 6 topics are covered:

The first contribution, by Elvé Moons from Alcatel Belgium, gives a good overview of DSL drivers. It compares the advantages and disadvantages of class-AB, D, G, and H drivers.

In the second contribution Thomas Ferianz from Infineon, Austria, describes class-AB bus drivers. The force of class-AB lies in high linearity and a reasonable efficiency at low crest factors.

Tim Piessens from the KU Leuven, Belgium, presents an efficient class-D driver using a Self-Oscillating Power Amplifier (SOPA).

The fourth paper by John Pierdomenico from Analog Devices Incorporation, USA, explains class-G and H drivers, that have an efficiency between that of class-AB and class–D drivers combined with the same high linearity as the class-AB drivers.

Two other types of computer busses are presented: Gerrit den Besten from Philips Research, Netherlands, explains the USB2 bus in the fifth paper.

Finally, Kimo Tam from Analog Devices, USA, presents problems and solutions of transceivers on computer backplanes in the sixth paper of this chapter.

Johan Huijsing

This page intentionally left blank

Looking to/for low power ADSL drivers in the DSLAM.

(DSLAM : digital subscriber line access multiplexer)

Elvé Moons
Fixed Network Division
Alcatel
Francis Wellesplein 1, B-2018 Antwerp, Belgium
elve.moons@alcatel.be
tel. +32 3 2407847

Abstract

An overview is given about the central office (CO) ADSL line driver. In the introduction it is explained there is still a high volume of components to be expected in the next years. Further on, some ADSL signal characteristics are highlighted in order to understand the power efficiency problem of the line driver, requirements are explained and critical performance issues considered. Topologies of existing drivers are described and include the class AB, G and H. Attention is give to switching line drivers like the class D, K and SOPA. Finally a combined architecture is described taking the benefit of high efficient switching together with high linear class AB operation.

1. Introduction

Looking back in time we see that in the 80's the main telecom growth driver was related to voice services, in the 90's the mobile service was taking off, today the major growth driver is surely the broadband service. Data traffic growth is increasing very rapidly due to the number of connections but also due to the increased number of services and applications. In the coming years, xDSL market and related business will continue to generate ever increasing revenues for telecom operators, OEM and semiconductor suppliers. Looking to forecasts of some analysts it is expected that the number of DSL connections will steadily increase the coming years with a growth rate of more than 25% per year, reaching over 100millions of installed lines by 2005.

Regarding the geographic mix, Asia and North America clearly had a fast start but looking to the future perspectives, Europe is fast catching up. It is expected that this acceleration in Europe will continue given that DSL technology is the more dominant broadband technology in this region. It is beyond any doubt that

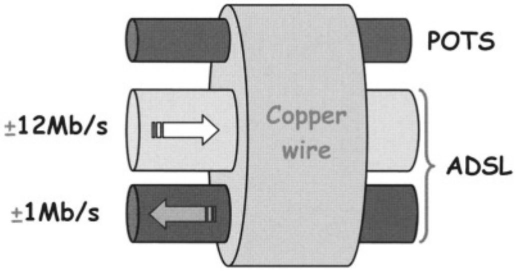
A. van Roermund et al. (eds.), Analog Circuit Design, 275-290.
© 2003 *Kluwer Academic Publishers. Printed in the Netherlands.*

the internet and its related service is part of every citizen's life, not even mentioning its economical and industrial impact. Clearly the availability of affordable high-speed networking and communications is of major concern for every community. Telecommunication companies have now developed high-speed wireline connections such as ADSL which today are in full deployment. Looking ahead on the need for speed generated by the offering of more extensive services such as multiple real-time video channels combined with Voice over Internet Protocol (VoIP) and more in general Voice over Broadband (VoB), more generally referred to as Voice over Packet (VoP), the development of the next generations of wireline communication such as VDSL (Very high speed Digital Subscriber Loop) is now ongoing.

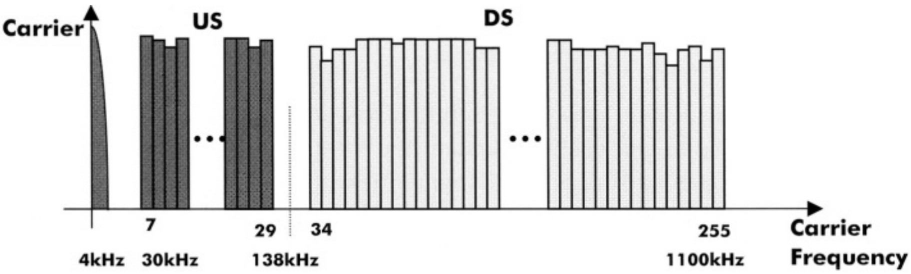
Looking for increased port density in the DSLAM (Digital Subscriber Line Interface Access Multiplexer), specially the power management and bill of materials (or integration level) becomes more and more important. Over 70% of the power per ADSL channel is related to the linedriver and its line termination. The ADSL discrete multitone (DMT) signal is a golden gate on copper cable but leaves the system with a power bottleneck in the line drivers because of the high crest factor. Class AB drivers are still in ADSL products but will soon be replaced with Class G, Class H, Class K and other high efficiency power amp concepts. We re-invent the driver architectures as we go back to the future and remember from the audio low power, high efficiency concepts. But the boundary conditions for DSL are more complicated. The extras are: line termination, echo cancelling and duplex signalling. These additional functional issues make the difference and prevent us from just copying the audio low power topologies. But especially the high peak to average ratio (PAR) and the high linearity requirements of the quadrature amplitude modulation (QAM) in the discrete multi tone (DMT) signal code make it hard to implement all these well known low power concepts. Three categories of topologies can be considered: linear, switching mode and combined architectures the last one using switching mode for the high efficiency power in the signal and a linear correction amplifier to achieve the high linearity requirement.

2. The ADSL signal: DMT, QAM, PAR and Power

ADSL signalling is done via the twisted pair between subscriber and central office (CO). It makes use of frequency division multiplexing (FDM) : voice signalling or POTS (plain old telephone service) happens below the 4kHz band, Upstream (US) ADSL signalling, from subscriber to CO, in the band 30kHz to 138kHz and downstream (DS) signalling from 138kHz to 1.1MHz. In this way full duplex transmission is achieved while voice and ADSL can happen simultaneously. Datarates of 12Mbit/s in DS direction and 1Mbit/s in US are possible for short loop lengths.

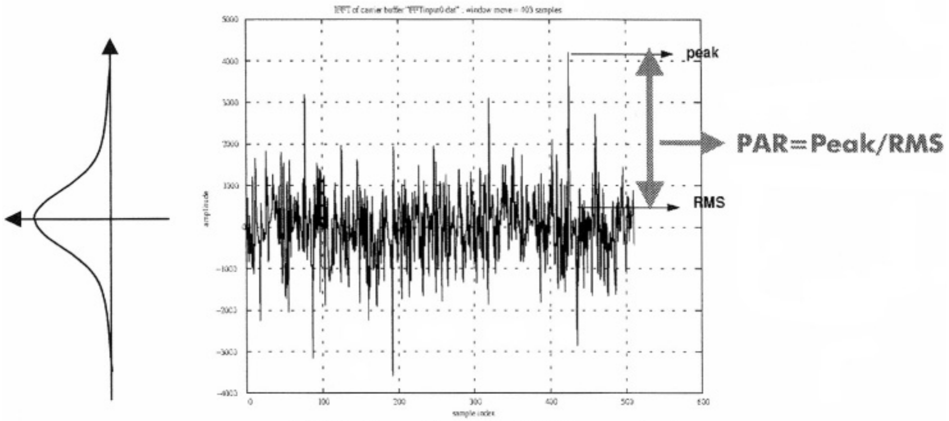

POTS
±12Mb/s
Copper wire
ADSL
±1Mb/s

The CO transmit or downstream (DS) signal of ADSL is built up by the addition of more than 200 tones in the band from 138kHz to 1.1MHZ and carrier spacing of 4.3125kHz.


Carrier
US
DS
7
29
34
255
4kHz
30kHz
138kHz
1100kHz
Carrier Frequency

ADSL spectrum

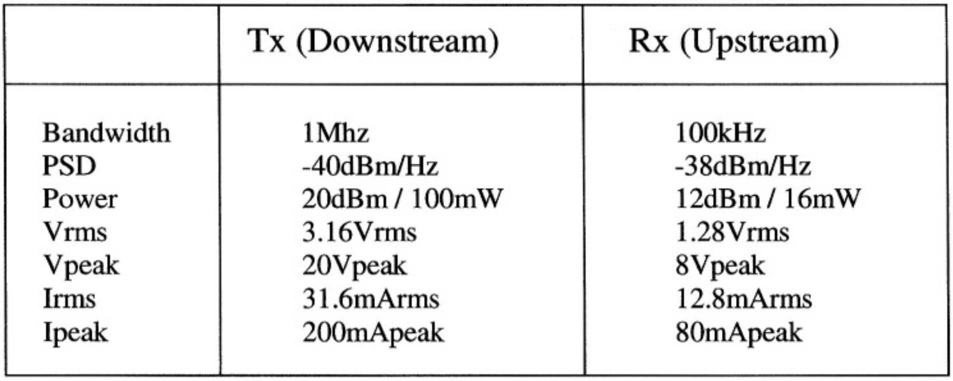
Each tone is modulated in amplitude and phase (QAM) and can contain up to 15 bits corresponding to 32768 constellation point in a phase amplitude diagram (per tone). The resulting signal in the time domain looks like white noise, it has a Gaussian amplitude distribution and is transmitted during 250μs. It is a signal with rather small average amplitude value but sometimes, when a larger number of tones have the same phase, can show a very high peak value.



ADSL time domain DMT symbol

In order to limit dynamic range of the signal path and as such power, signal to noise ratio and cost (circuit area), the peak to average value (PAR) or Crestfactor is clamped in the system (digital domain). This result in a disturbance in the frequency band and can cause bit errors due to the reduced signal to noise ratio. The probability of occurrence of such bit errors is determined by the chosen PAR and as such results in a certain bit error rate (BER). In ADSL a PAR of 6.4 (or 16dB) is chosen to achieve a BER lower than 10^{-7}.

Taking into account the power spectral density of -40dBm/Hz, for the downstream signal one can easily calculate the total power level of 20dBm or 100mW. Knowing the subscriber line is equivalent to a 100Ω impedance and a PAR of 16dB this result in a rms signal of 3.16Vrms and 20Vpeak (or 40Vppd) on the line. The following table summarises the main characteristics for both the US and DS signal on the line.



	Tx (Downstream)	Rx (Upstream)
Bandwidth	1Mhz	100kHz
PSD	-40dBm/Hz	-38dBm/Hz
Power	20dBm / 100mW	12dBm / 16mW
Vrms	3.16Vrms	1.28Vrms
Vpeak	20Vpeak	8Vpeak
Irms	31.6mArms	12.8mArms
Ipeak	200mApeak	80mApeak

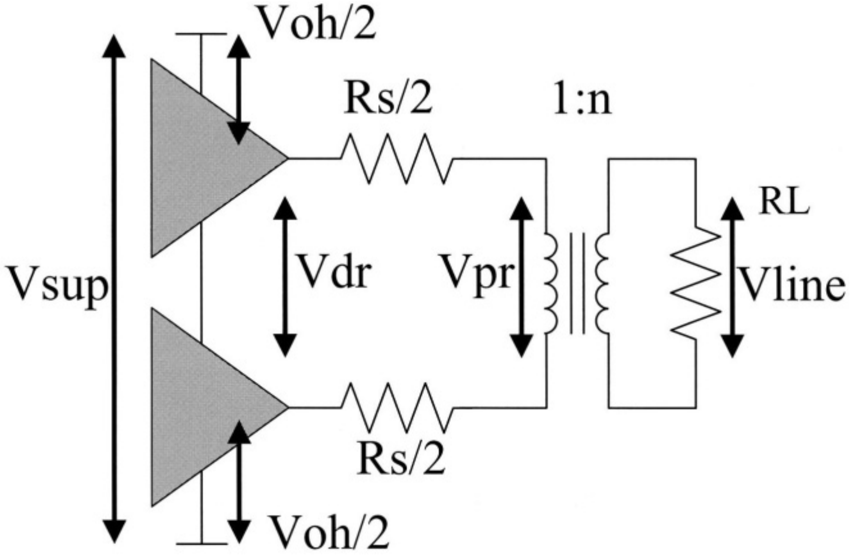
ADSL downstream and upstream signal characteristics

3. CO line driver efficiency and requirements

The driver should generate via a transformer the transmit signal on the line, provide voltage gain and realise the matched line impedance. Main requirements are low noise, low distortion, high slewrate and good efficiency.

The gain requirement is just from the fact that the driver receives its input signal from a low voltage analog front end IC usually limited to a amplitude level of 2Vp. Changing the transformer turns ratio influences the gain requirement and allows a flexible choice of supply voltage.

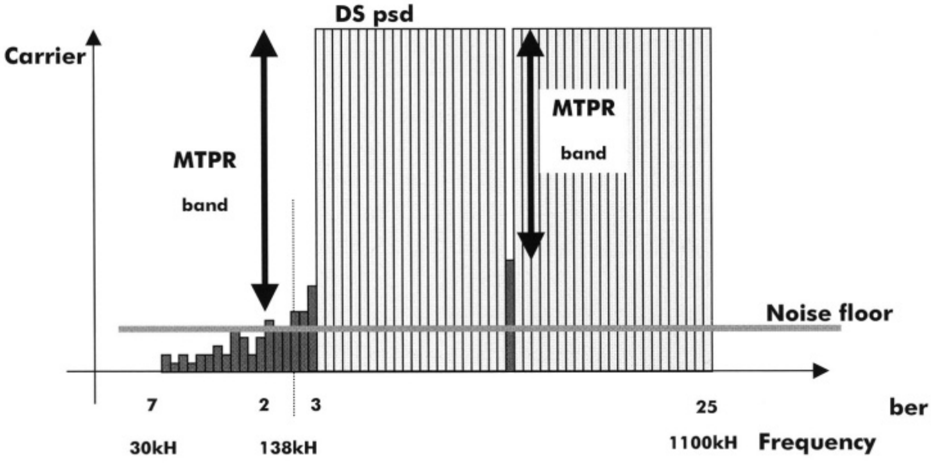
To realise the 100Ω output impedance, one can just add a pair of series resistors in the transmit path while at the same time one is able to receive the US signal by measuring across this resistors. This however means that the same power (100mW) is dissipated in the termination and that maximum efficiency is already reduced to 50%. Further more it also means that the driver output signal is twice the signal at the line and that supply voltages have to scale accordingly. Using a transformer with turns ratio doesn't change as such anything for the power consumption.



Line Driver, Termination and Transformer (terminology)

Looking to a class AB type driver, a maximum efficiency of 78% can be achieved for a sine wave signal without taking into account quiescent current (Iq) nor overhead voltage (Voh) required by the output transistors. For a DMT signal however, due to its high crestfactor, this maximum efficiency reduces to about 20%. Also taking into account overhead voltage and quiescent current only 7% efficiency can be reached resulting in a power consumption of 1.5W or 1.3W dissipation in the line driver. To show this with a practical example one can choose for a transfo turnsratio of n=2, a crestfactor CF=6.4 and a overheadvoltage Voh=4V. This result in a supply voltage Vsup=2*Vrms*CF/n+Voh=24V. The mean current for a DMT signal is about 80% of its rms-value. Added with a quiescent current Iq=12mA for the driver results in a mean power supply current of about 60mA, and so ending up with a 1.5W power consumption.

Besides to fulfil the signal characteristics of the transmitted signal, most of the critical requirements of the driver are imposed by the US received signal. In order to be able to do a 15bit QAM per tone in transmit, a signal to noise and distortion ratio of about 65dB is required. This can be characterised by the parameter MTPR: the missing tone power ratio (MTPR). Within the DS band a full transmission happens except for one tone, the depth of the missing tone is measured and represents the MTPR. More important however is the influence towards the US band. Injection of noise and distortion, due to intermodulation, can happen within the driver disturbing or a least limiting the sensitivity of the received signal. Although a hybrid function is reducing the echo signal from driver to receiver, in practice this suppression is limited to 10 to 20dB. If the driver has to achieve the background noise level of -140dBm/Hz of the line, the driver's noise and distortion levels in the US band should not exceed -120dBm/Hz. This means a transmit psd to receive noise and distortion ratio of more than 80dB.



Missing Tone Power Ratio (MTPR)

4. The POWER issue

In the further evolution of the DSLAM density increase and the drive to put more lines per board (for cost reasons), the thermal management becomes a big issue. More than 70% of the power dissipation of an ADSL channel is related to the driver and lack of good driver efficiency prevents the increase of line terminations per board.

On the other hand we are interested to satisfy the NEBS power requirements to allow us to fill the DSLAM in all cases. The NEBS requirement limits the power dissipation per square meter floorspace. Translating to the line termination board in the DSLAM, this puts a limit on the total dissipation per board. While moving from 12 and 24 lines per board to 48 or even more, this really becomes a thermal problem. In the table below one can see the maximum allowed driver dissipation for a $500cm^2$ board in function of the number of terminals per board. In order to reach the NEBS requirement for a 48line board the power dissipation of the line driver should be below 500mW (see table below).

# lines/board	12	24	48	72	96
Pdiss. Driver (mW)	2310	1080	465	260	160

Line Driver dissipation related to NEBS requirement

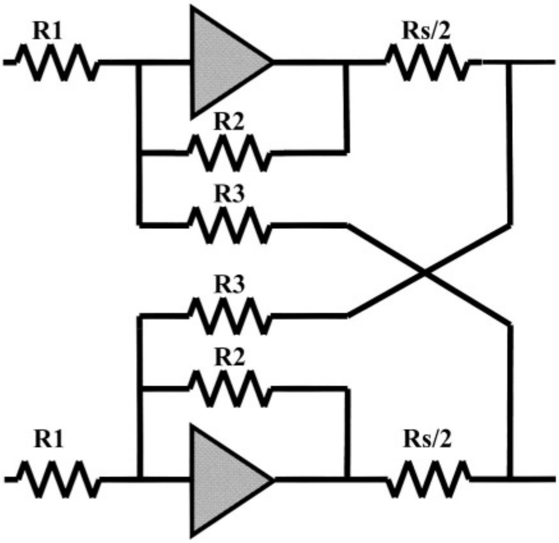
5. Active Back Termination (ABT)

Impedance synthesis or active back termination is a first means to lower the power dissipation. In stead of putting the real termination resistor in series with the line driver, only a fraction is placed while an active feedback loop amplifies the value to generate the required Impedance. For resistive impedance synthesis, the implementation simplifies to the schematic shown below in which feedback resistors R3 are added.

$$Gain = \frac{R3.R2}{(R3 - R2).R1}$$

$$Gain = \frac{R3.R2}{(R3 - R2).R1}$$

$$ABT = \frac{Rs + Rout}{Rout}$$



Active Back Termination (ABT)

The power is reduced in two ways. Because of the reduction of the resistor, the power dissipation, relative to the 100mW dissipation in the line, is reduced proportionally. From the other hand also the signal level at the driver output is reduced and so the required power supply voltage. Using the same conditions as above with an ABT-factor=1.25, supply voltage reduces to 16.5V and power consumption to 1W or efficiency to 10%. Further decrease in the ABT factor becomes difficult for stability reasons and receiver sensitivity (RX attenuation and related noise requirements).

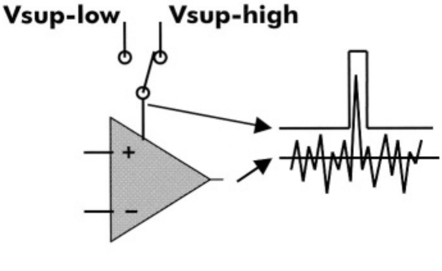
Optimisation of ABT, quiescent current, overhead voltage and transfo turns ratio results in today's class AB drivers reaching 750mW power consumption (see ISSCC 2002). Not to forget are the solutions with reduced turns ratio where the CO battery can be used directly as supply for the driver and in this way avoiding the loss in the DC/DC converter on the ADLS LT-board while at the same time improving the receive sensitivity.

It is however clear that the NEBS requirement can not be met with the class AB approach. Therefore other topologies and architectures have to be considered, some of them known out of the audio world.

6. Driver topologies and architectures

Class G

A first class of driver to think about is the class G one. In this driver 2 pairs of supply voltage exist, and depending on the value of the required output, the current of the output stage is coming out of a low or high supply. The DMT signal is very suitable for this kind of principle because most of the time the signal amplitude is limited while only during a very small time interval a high voltage output has to be delivered.



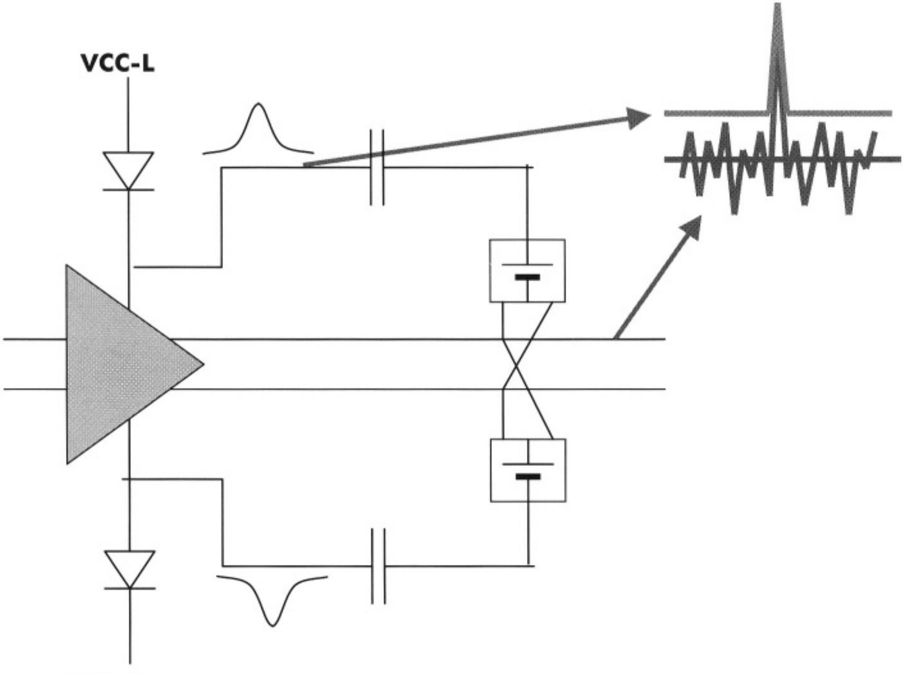
Class G amplifer

A drawback of the class G driver is the additional pair of supplies. Further more it is very difficult to achieve the severe linearity specifications due to the hard switching and related crossover points. The switch control can be done from the digital part (digital prediction), from the analog input or directly driven from the output. Techniques for having smoother transitions during switching can/must be applied to limit their impact towards the output. The complexity is within the amplifier providing a good power supply rejection.

Related to the DMT signal and optimal power consumption results in a 2.5 to 3 ratio of high to low voltage supply. The power consumption can be as low as 600mW.

Class H

The class H amplifier is similar in power consumption compared to the class G approach. In stead of hard switching, the supply voltage is progressively following the output signal. This avoids the need for good power supply rejection but off course requires additional circuitry for generating the driver supply. The supply generation can be done for instance by a capacitive boosting of the supply voltage applied. Power consumption compared to class G remains however equal but only 1 pair of supplies is required. Although only 1 pair of supplies is needed, the number of capacitors on the board is not reduced and one should be aware of signals exceeding the applied supply voltage.



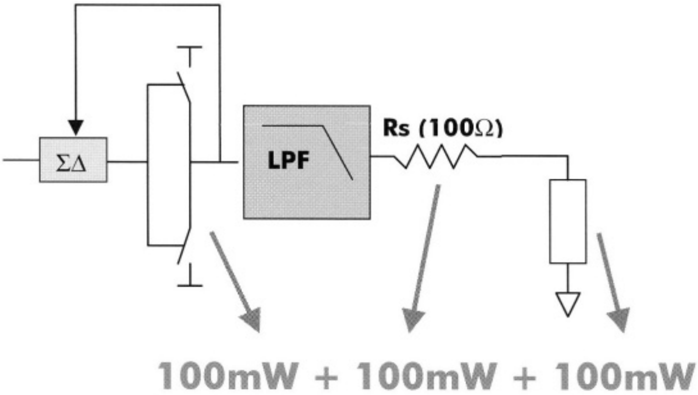
Class H amplifier

Switching line drivers (SLD)

Class D

Switching line drivers are tempting us for the well known high power efficiency of switching mode systems in general. Audio amplifiers have used switching techniques, switching mode voltage regulators also, but again the extras: line termination, echo cancelling and duplex signalling are giving a new challenge for analog designers to extend the use of "power DAC " and **"ΣΔ"** to ADSL line drivers.

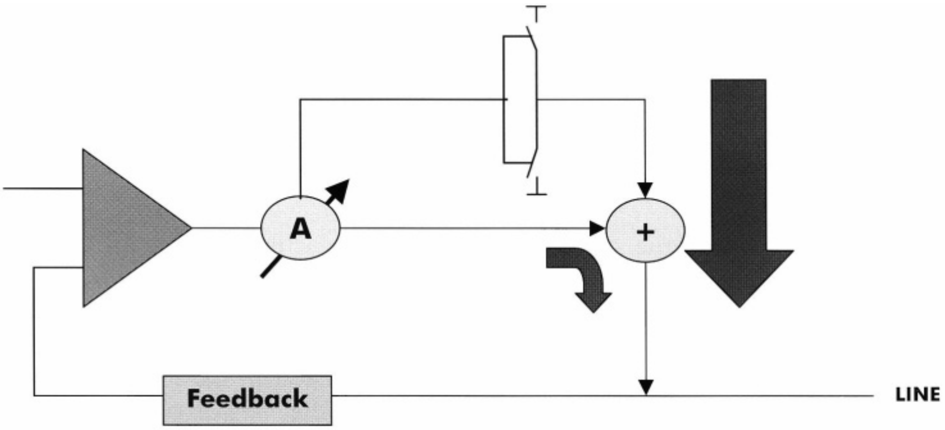
Power consumption of the switching line driver can achieve the 300mW limit. To achieve the high linearity figures of 80dB MTPR, the order the of the **ΣΔ** must go up to **5th** or **6th** for single bit while on the other hand the multibit approach requires additional supply voltages. Active back termination can not as such be applied due to the digital to analog conversion in the feedback path. Further more some passive filtering is required so that boardspace usage and cost might become an issue again. Also finding a suitable technology combining high voltage and speed is not so obvious and the problems will surely not reduce if thinking about higher bandwidths.



Class D amplifier

Class K

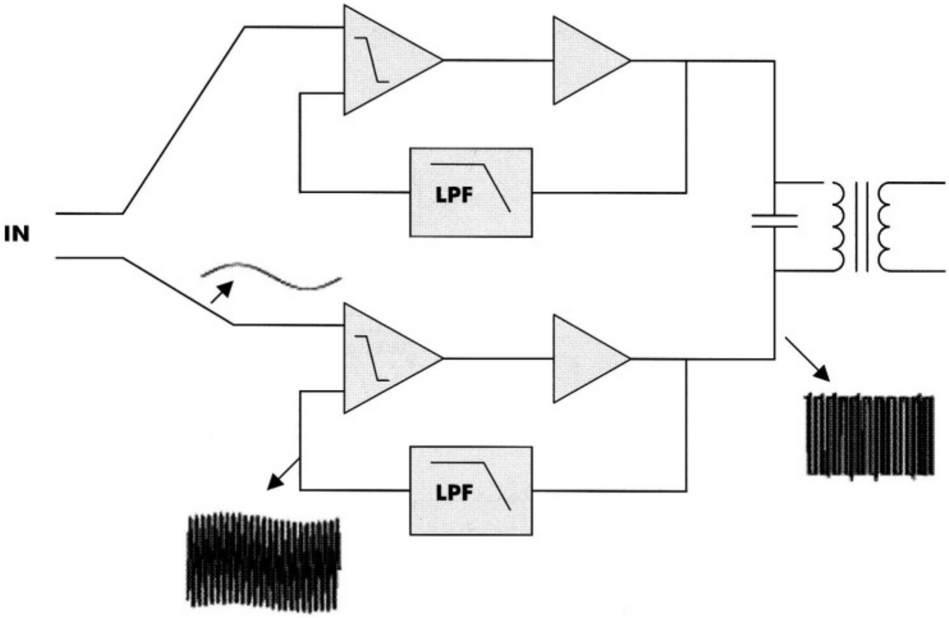
Another type of switching line driver is the class K. In this topology the main current is delivered via high efficiency switching while control is done via linear feedback. This at least reduces the filler requirements and has been applied in the audio world, however must still be proven for ADSL applications.



Class K amplifier

SOP A : Self Oscillating Power Amplifier

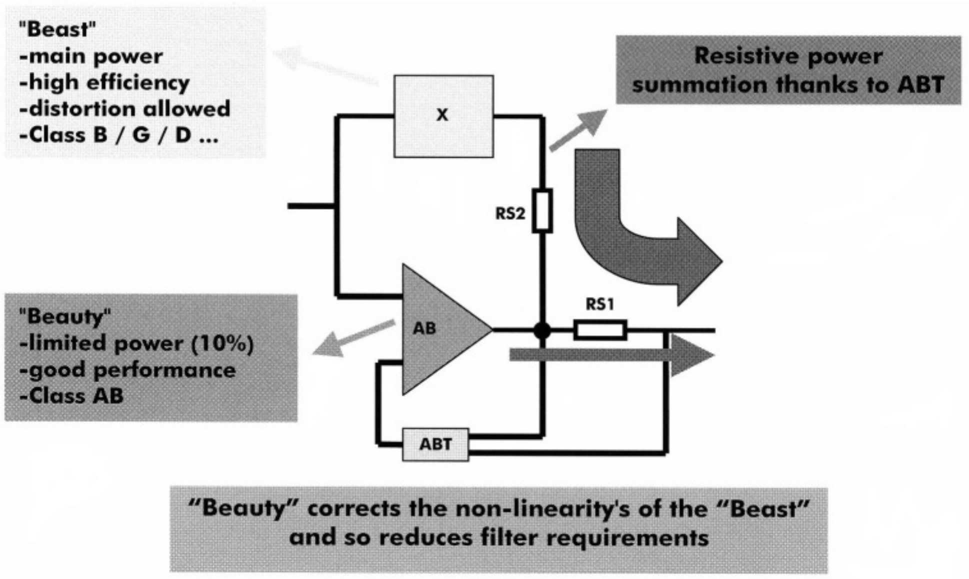
The SOPA structure is also based on a switching operation mode. The frequency is however due to a self oscillating loop without any clock being applied. Due to the symmetric and matched topology the both oscillations are pulled to each other and as such act like a common mode signal being rejected in the transformer. Due to the self oscillation a linearisation happens and reduces the need for filtering. Here also research has been done but some extra work for ADSL CO drivers is still required.



Self Oscillating Power Amplifier (SOPA)

Combined structures

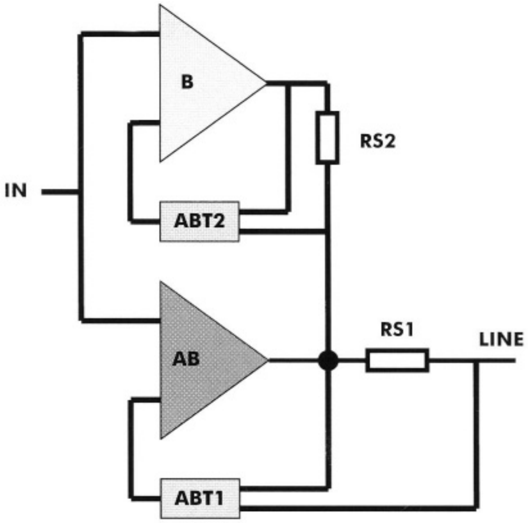
The concept is based on the idea to combine the advantages of linear drivers and switched mode drivers. Why not deliver the main power via switching techniques while monitoring and correcting the output via a linear amplifier, similar to the class K structure. This concept is given the name “Beauty and the Beast”.


"Beast"
-main power
-high efficiency
-distortion allowed
-Class B / G / D ...
X
Resistive power summation thanks to ABT
RS2
RS1
"Beauty"
-limited power (10%)
-good performance
-Class AB
AB
ABT
"Beauty" corrects the non-linearity's of the "Beast" and so reduces filter requirements

Combined Class Amplifier ("Beauty and the Beast")

The "Beast" or power driver is delivering most of the power with high efficiency but moderate linearity. It can be a class B , G or switching amplifier in which quiescent current or losses are minimised. The "Beauty" or correction amplifier can than deliver a limited power to linearise the signal before it reaches the line. This amplifier can use its maximum capability to do this and is offered to do so due to the limited power it should deliver. It can be a high bandwidth, low noise, low power class AB amplifier. The "Beauty" in this way corrects the deviations imposed from the "Beast" while limiting itself to low power operation. It can even reduce the filter requirements needed when implementing a switched line driver. It is expected that power consumption will be below 550mW. It is thanks to the use of active back termination that the power summation can be done in a easy resistive way (see figure below in which as example a class B-AB is chosen).

A first design condition is that output impedance on the line is equal for bot paths. For the class AB amplifier this is simply the $RS1 * \alpha_l$, where α_l includes the active back termination. For the B amplifier the output impedance at the line includes RS1 and only the remaining part needs to be synthesised with a local active back termination. When also the gain from input to the line is the same than both amplifiers want to generate the same voltage at the class AB output, however the B amplifier performs this without feedback and as such delivers the main power. The deviations from the wanted signal are measured via the local feedback around the class AB amplifier and corrected if needed.



Design conditions
Gain : $(LINE/IN)_1 = (LINE/IN)_2$
output impedance : $Rout = \alpha_1.RS1 = \alpha_2.RS2 + RS1 = Rline$
impedance amplification factor $\alpha = 1/(1-ABT)$

Class B-AB Amplifier

7. Conclusion

A lot of effort has already been spend on the realisation of low power ADSL line drivers for the CO side. Coming from the class AB stage with a 700mW limitation, gradually new topologies like class G and H are introduced, limiting the power consumption to 600mW. Further research and development is however needed to further reduce this power consumption in order to allow further density increase in the DSLAM. Switching line drivers can reach this however filter requirements and related boardspace must be careful watched to finally reach the goal. A new concept with a combined topology has been presented which can lower the power consumption in the mean time without the need for to much filtering requirements.

References

1) R. van der ZEE and E. van Tuijl, “A power- efficient audio amplifier combining switching and linear techniques.”, IEEE JSSC july 1999, p. 985-989.

2) J. Pierdomenico, Scott Wurcer, Bob Day, “A 744mW Adaptive Supply Full Rate CO Driver.”ISSCC 2002, p. 320.

3) Faramarz Sabouri, Reza Shariatdoust, “A 740 mW ADSL Line Driver for central Office with 75dB MTPR.”, ISSCC 2002, p 322.

4) Nam-Sung Jung et al., “A new high efficiency and super fidelity Analog Audio Amplifier with the aid of digital switching amplifier : Class K amplifier.”, IEEE p. 457 to 461.

6) Domenico Rossi, “Scalable ‘High voltage’ integrated circuit design for XDSL type of applications”, AACD workshop 2001.

7) Tim Piessens , Michiel Steyaert, “SOPA : A high-efficiency line driver in 0.35mm CMOS using a self-oscillating power amplifier.”, ISSCC 2001, p. 306,307.

CLASS-AB LOW-DISTORTION DRIVERS FOR ADSL

Thomas Ferianz
Infineon Technologies
Siemensstrasse 2
A-9500 Villach, Austria

Abstract

Advantages and disadvantages of different line driver types are investigated and the efficiency of class-AB drivers is calculated. In addition, three class-AB line drivers implemented in 0.6µ standard CMOS, high voltage (HV) BiCMOS and complementary bipolar (CB) processes, respectively, are described.

1. Introduction

Driven by the increasing port density in an ADSL Central Office (CO) in the past several years, the power consumption of the low-distortion line driver becomes a key parameter for system design. Due to the high Crest Factor (CF) of Discrete Multi Tone (DMT) signals, the line driver has to deal with very high signal peaks of up to 40Vpp on the line, thus requiring a high supply voltage. This leads to low line driver efficiency. To closely approach the theoretical efficiency limit, the voltage drop and the quiescent current of class-AB drivers have been reduced consequently with every new product generation.

In Section 2, the strengths and weaknesses of current-feedback and voltage-feedback amplifiers are discussed.

In Sections 3-5, a detailed description of three different class-AB ADSL CO driver circuits is given. Implementations in 0.6µ standard CMOS, high voltage BiCMOS and complementary bipolar processes are addressed.

Section 6 covers the calculation of class-AB driver efficiency for a DMT signal with Gaussian amplitude distribution.

2. Current-Feedback versus Voltage-Feedback

In the 1980s, the current-feedback topology shown in Fig. 1 has emerged as the dominant choice for broadband amplifier designs [1]. This type of amplifier can achieve slew rates of several thousand V/µs, since the amount of

A. van Roermund et al. (eds.), Analog Circuit Design, 291-308.
© 2003 *Kluwer Academic Publishers. Printed in the Netherlands.*

slew current is proportional to the input voltage. Any large input signal step momentarily causes a large voltage across the feedback resistor Rf2. The available current to charge the total compensation capacitance Ct is proportional to this voltage. Due to the absence of first-order slew rate limiting, current-feedback amplifiers are well suited for low-power circuits. However, in real implementations second-order effects limit the slew rate.

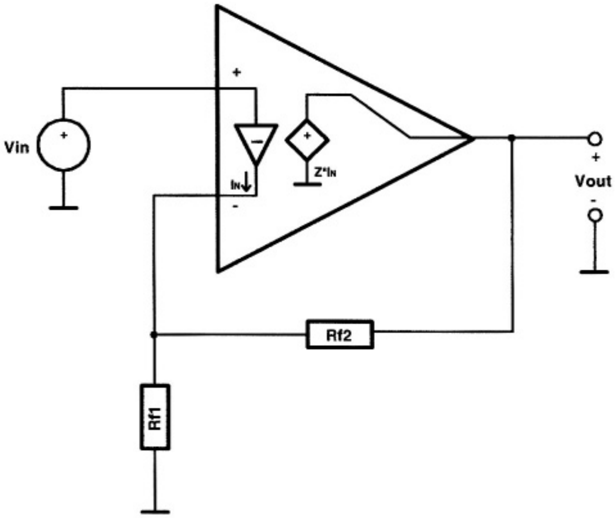


Fig.1. Non-inverting current feedback amplifier

Another advantage of the current-feedback amplifier can be seen from the equation determining the closed-loop gain of a non-inverting amplifier structure with feedback resistors Rfl and Rf2:

$$A_{cl} = \left(1+\frac{R_{f2}}{R_{f1}}\right)\cdot\left(\frac{1}{1+j\cdot\omega\cdot R_{f2}\cdot C_t}\right) \tag{1}$$

The closed-loop pole is set by the total compensation capacitance Ct and the feedback resistor Rf2. Consequently Rf2 can be used to select the –3dB bandwidth, whereas Rf1 determines gain [2].

A very popular implementation of a single-stage current-feedback amplifier is shown in Fig. 2. In this simple topology Wilson current mirrors perform the level shift to the output stage and inject the slew current to the compensation capacitor Cc. If the achievable open-loop gain, determined by the transimpedance of the circuit, is not sufficient, another gain stage may be added.

Nevertheless, current-feedback amplifiers generally suffer from high input offset current, thermal feedback errors, high noise and poor CMRR [3].

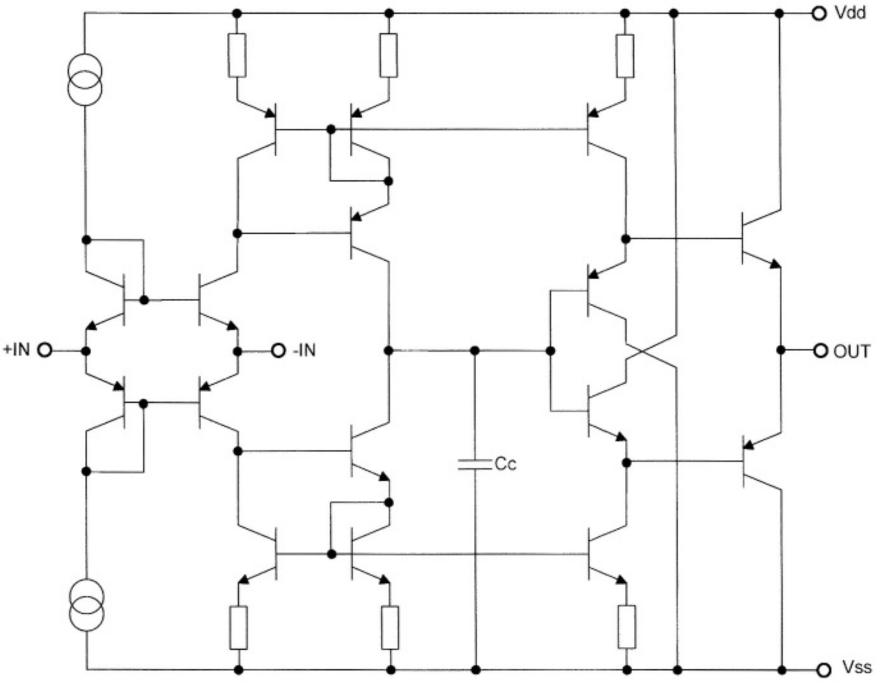


Fig.2. Single stage current-feedback amplifier

In the past decade, voltage-feedback amplifiers have reemerged [1]. Compared to their current-feedback counterparts, they offer several features like low noise at low closed-loop gains or the ability to build inverting integrators for active filter applications and fully differential, inverting line drivers. For active filter applications and as a preamplifier in line driver circuits, the well-known single-stage folded-cascode design shown in Fig. 3 is the best choice. Both noise characteristic and small signal bandwidth of this amplifier are excellent. To overcome the limited slew rate and thereby small full-power bandwidth of voltage-feedback amplifiers with class-A input stage biasing, class-AB input stages can be used. One possibility to implement this stage into an operational amplifier is to replace the unsymmetrical input-stage in Fig. 2 with the symmetrical one shown in Fig. 4. Due to the increased number of devices involved in the class-AB biasing, the voltage noise of the input stage rises [1].

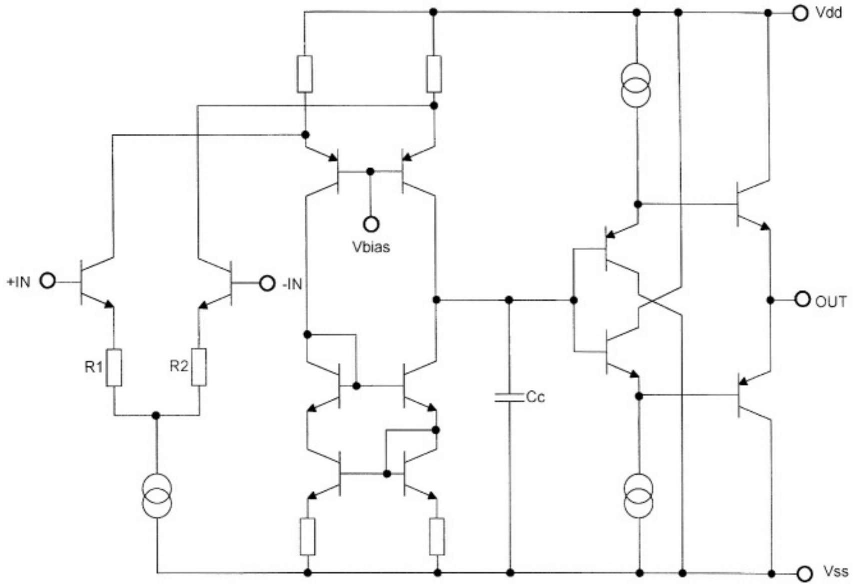

Vdd
Vbias
+IN
-IN
R1
R2
Cc
OUT
Vss

Fig.3. Folded-cascode amplifier with class-AB output-stage

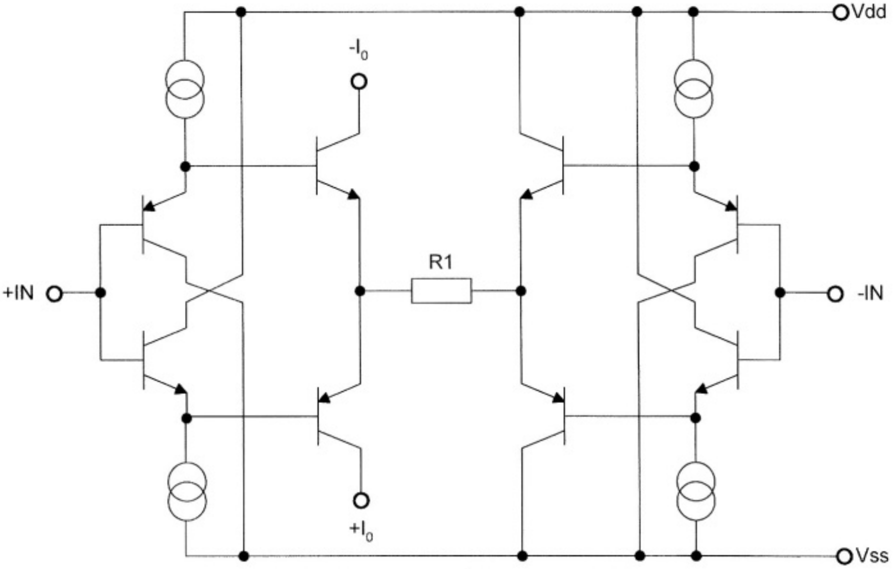

Vdd
$-I_0$
+IN
R1
-IN
$+I_0$
Vss

Fig.4. A symmetrical class-AB input-stage

3. A CO-ADSL line driver with integrated supply switch

To achieve low power consumption, the line driver shown in Fig. 5 is configured to synthesize 83% of the output impedance by means of an internal feedback loop established with resistors R3 and R3'. The radical reduction of series impedance reduces the supply voltage needed at a given line power level.

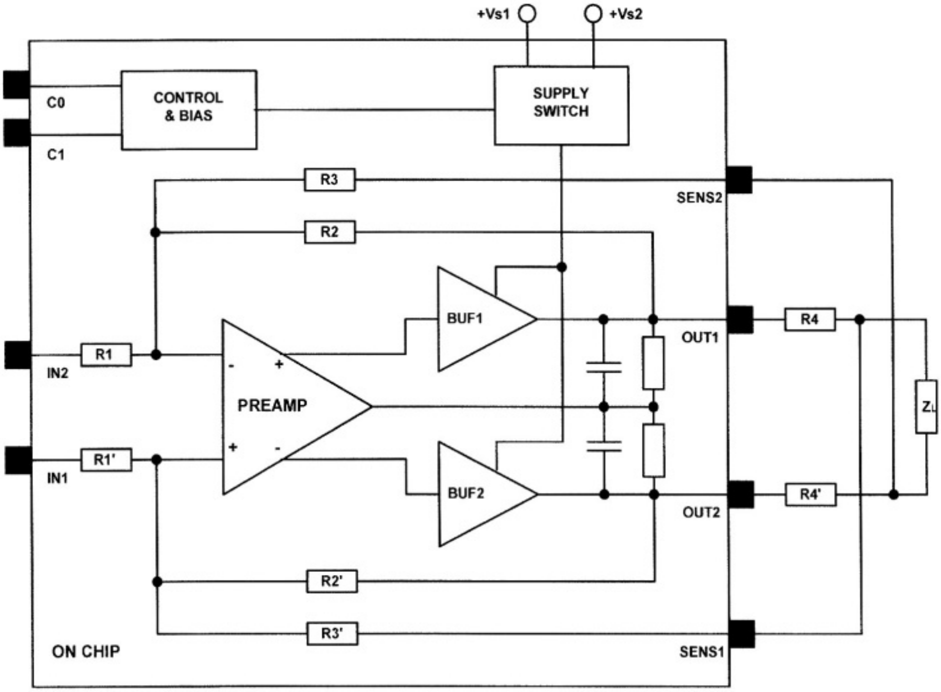

+Vs1
+Vs2
C0
C1
CONTROL & BIAS
SUPPLY SWITCH
R3
SENS2
R2
BUF1
OUT1
R4
R1
IN2
-
+
PREAMP
ZL
R1'
IN1
+
-
BUF2
OUT2
R4'
R2'
R3'
ON CHIP
SENS1

Fig.5. Block diagram of ADSL line driver with supply switch

The output resistance can be calculated to be

$$R_O = m \cdot \left(R_4 + R_4'\right). \tag{2}$$

The synthesis factor m is determined by the values of R3 and R2,

$$m = \frac{R_3}{R_3 - R_2} \tag{3}$$

$$R_2, R_2', R_3, R_3' >> R_4, R_4'$$

To further decrease power consumption, a supply switch is integrated. It allows to supply the output stage with a lower voltage if the transmit power is reduced at short or very long loops. The voltage for both supplies can vary between 5V and 21V.

For the two output buffers single-stage current-feedback amplifiers are used. They offer the required linearity and bandwidth, while consuming only 1.5mA of quiescent current. Fig. 6 shows the schematic diagram of the output buffer. As in any current-feedback amplifier, the input stage Q10, Q11, Q13 and Q14 is a unity-gain buffer forcing the inverting input at the emitters of Q13 and Q14 to follow the noninverting input.

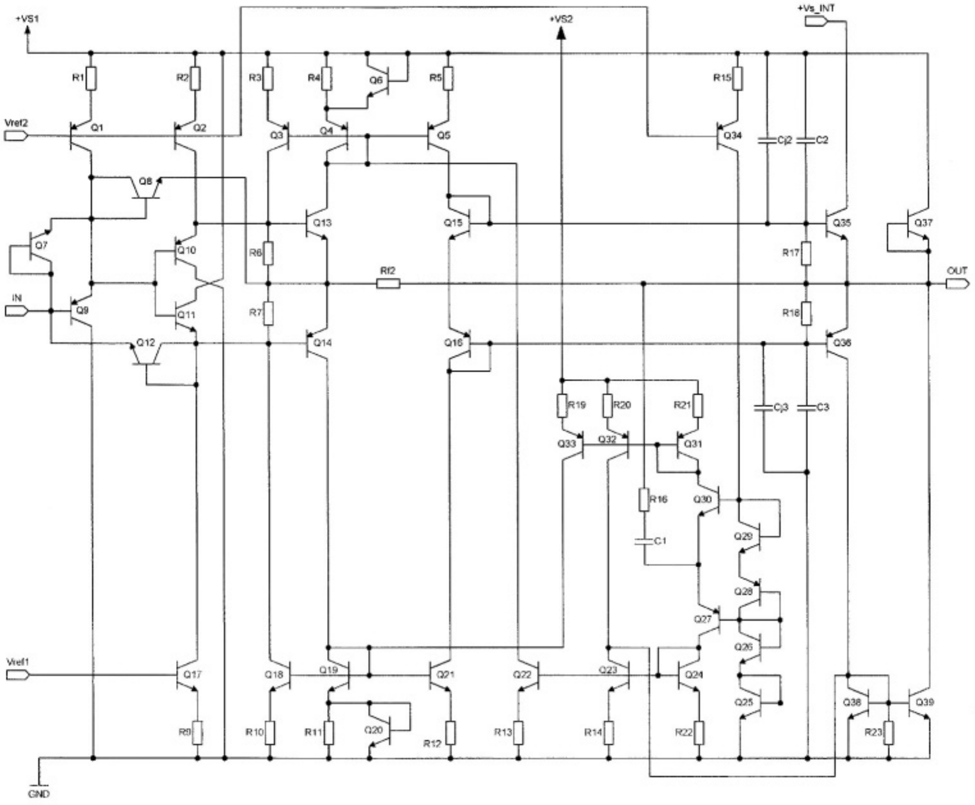

+VS1
+VS2
+Vs_INT
Vref2
IN
OUT
Vref1
GND
R1
R2
R3
R4
R5
R6
R7
Rf2
R9
R10
R11
R12
R13
R14
R15
R16
R17
R18
R19
R20
R21
R22
R23
C1
Cj2
C2
Cj3
C3
Q1
Q2
Q3
Q4
Q5
Q6
Q7
Q8
Q9
Q10
Q11
Q12
Q13
Q14
Q15
Q16
Q17
Q18
Q19
Q20
Q21
Q22
Q23
Q24
Q25
Q26
Q27
Q28
Q29
Q30
Q31
Q32
Q33
Q34
Q35
Q36
Q37
Q38
Q39

Fig.6. Schematic diagram of chosen buffer circuit

Follower Q9 further increases the input impedance of the buffer to reduce the load at the high impedance output node of the pre-amplifier. Any input voltage now causes current flow in or out of the inverting input. The amount of current depends on the value of feedback resistor Rf2 coupling the output of the amplifier to its inverting input. The signal currents are fed to the high impedance node by means of current mirrors Q4, Q5 and Q19, Q21 to charge and discharge the compensation capacitors C2, Cj2, C3, Cj3 and to provide the base current required by the output transistors Q35 and Q36. The mirrored signal current multiplied by the impedance seen at the high impedance node (bases of output transistors Q35 and Q36) sets the open-loop gain. Thus the output impedance of the mirrors and the reflected load impedance directly affect loop gain. From equation (1) the closed-loop pole that determines the -3dB bandwidth of the buffer, is given by the total capacitance Ct=C2+C3+Cj2+Cj3 and the feedback resistor Rf2. The input stage is biased by reference voltages Vref1 and Vref2 connected to the bases of Q2 and Q17. Resistors R2 and R9 determine the quiescent current for the input-stage. The current through Q10 and Q11 defines the base bias voltage for the transistor pair Q13 and Q14, while diodes Q15 and Q16 are used to bias the output transistors pair Q35 and Q36, respectively. Transistor pair Q38, Q39 guarantees adequate output current drive even for lowest beta of pnp transistor Q36. Load current flowing through Q36 is reflected through the current mirror action of Q38 and Q39. To minimize the even harmonics, the total current gain from the base of Q36 to the signal output must be equal to the beta of Q35. Thus the emitter area ratio m of transistor pair Q38 and Q39 is adjusted according to $m = \beta_{Q35} / \beta_{Q36}$. Q6 and Q20 are short circuit protection transistors. They limit the maximum amount of base current available for the output transistors Q38 and Q39. Q7, Q8 and Q12 clamp the buffer input to the output. They are off under normal operating conditions. In case of line overvoltage, diode Q37 clamps the buffer output to the positive supply. The buffer circuit is able to deliver an output current of 250mA; a thermal shutdown protection guarantees that the junction temperature does not exceed the maximum allowed value.

One remarkable detail of this buffer is a dynamical bias method for some current mirrors within the circuit to avoid high quiescent current and thus increase efficiency. For low frequency signals the output stage is biased in the AB region at a level high enough to eliminate crossover distortion. At high signal frequencies, fast slewing appears and current mirrors Q19-Q21 and Q38-Q39 are forced to shut off during the falling period of the positive signal half wave due to the action of parasitic base-collector capacitances Cjc of Q14, Q21, and Q36, Q39. Similarly, Cjc of Q5 and Q13 would shut off current mirror Q4-Q5 during the rising period of the negative signal half wave. This limitation can be overcome using the circuit based on transistors Q22-Q33. If slewing appears at the output node, a capacitive current is fed through the capacitor Cl and the resistor R16 into the virtual ground node at the emitters of Q27 and Q30. This current is recirculated through different current mirrors and fed back into the input nodes of current mirrors Q4-Q5, Q19-Q21 and Q38-Q39. By applying the required charge current for the parasitic capacitances, shut off in the mirrors and the corresponding crossover distortion can be avoided. R16 prevents capacitive loading of the output stage due to capacitor C1.

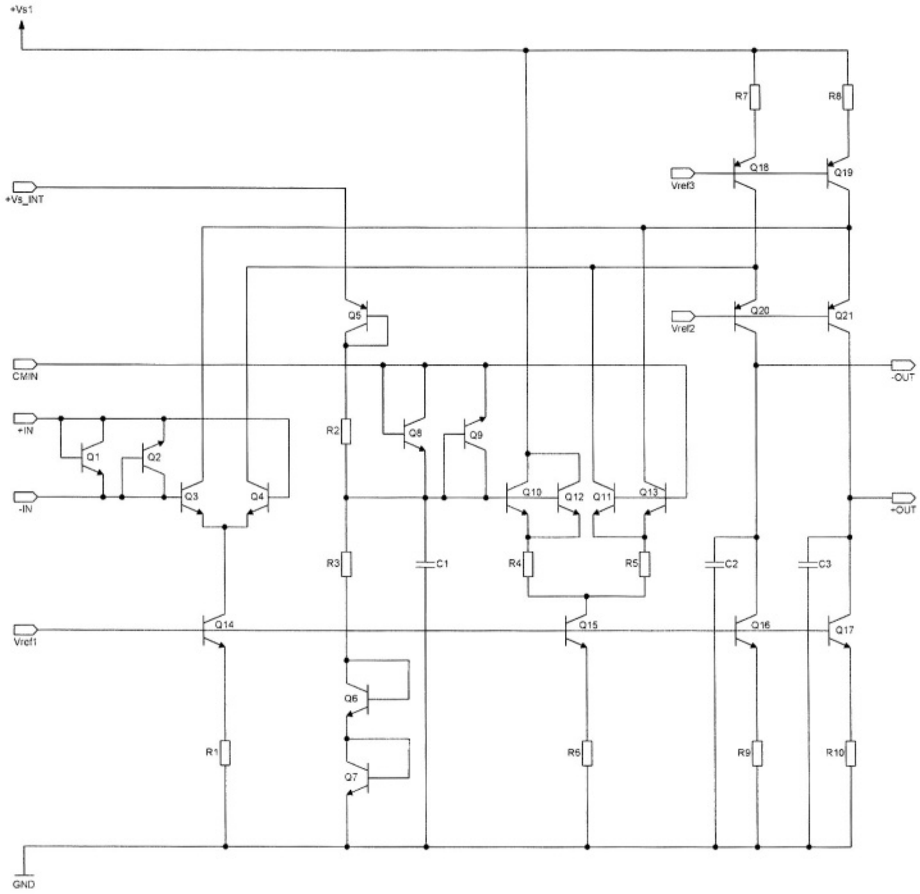

+Vs1
R7
R8
+Vs_INT
Vref3
Q18
Q19
Q5
Vref2
Q20
Q21
CMIN
-OUT
+IN
Q1
Q2
R2
Q8
Q9
Q10
Q12
Q11
Q13
-IN
Q3
Q4
+OUT
R3
C1
R4
R5
C2
C3
Vref1
Q14
Q15
Q16
Q17
Q6
Q7
R1
R6
R9
R10
GND

Fig.7. Schematic diagram of pre-amplifier

To achieve a Missing Tone Power Ratio (MTPR) of –75dBc, the preamplifier has to provide additional loop gain within the whole signal-bandwidth. The chosen fully differential folded-cascode structure shown in Fig. 7 is well suited to fulfil this requirement. While differential pair Q3, Q4 provides the differential gain, transistors Q10-Q13 are used to regulate the common mode output voltage.

Measured transmit performance is shown in Fig. 8 for a 100kHz/9Vp-p differential sine-wave signal. It clearly illustrates that the harmonics are in the -80-dBc region.

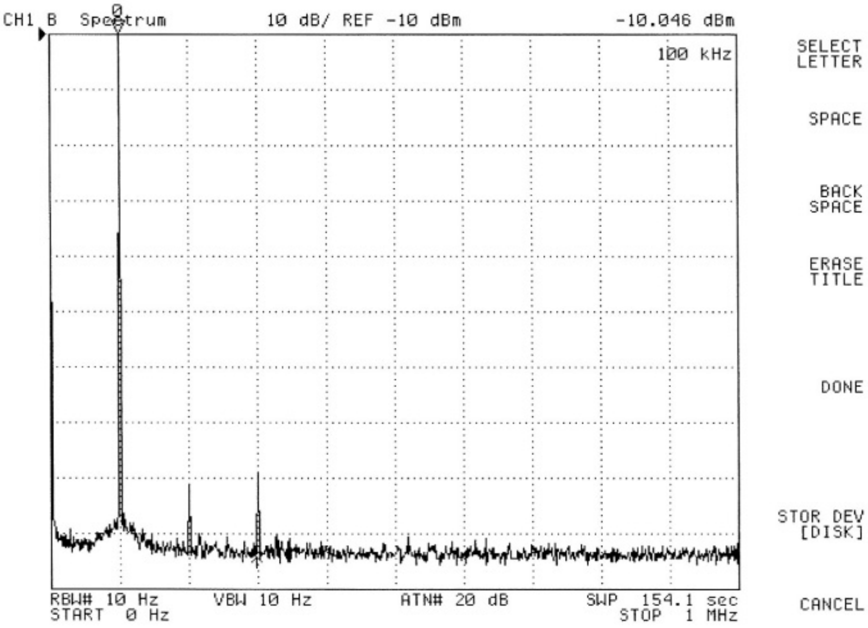


Fig.8. Spectrum plot of 100kHz/9Vp-p differential sine wave

The line driver uses the complementary bipolar part of a 0.6-µm BiCMOS double poly triple metal (DPTM) process with 1.5-GHz pnp and 4-GHz npn devices.

4. A CMOS ADSL line driver with current-mode output stage

The line driver in Fig. 9 is able to deliver 15Vpp into a 25 Ohm load. This is sufficient to transmit ADSL signals with a total power of 19.8dBm and a CF value of 4.5 into 100Ω via a 2:1 transformer. The line driver manages to meet this performance with an output stage fed between 0 and 10.5V and a total power consumption of 750mW.

A 0.6µm standard CMOS process is used for fabrication. To allow 11V supply voltage for the output stage, a NMOS transistor with extended drain is required. It can be implemented without violating the design rules of the standard process.

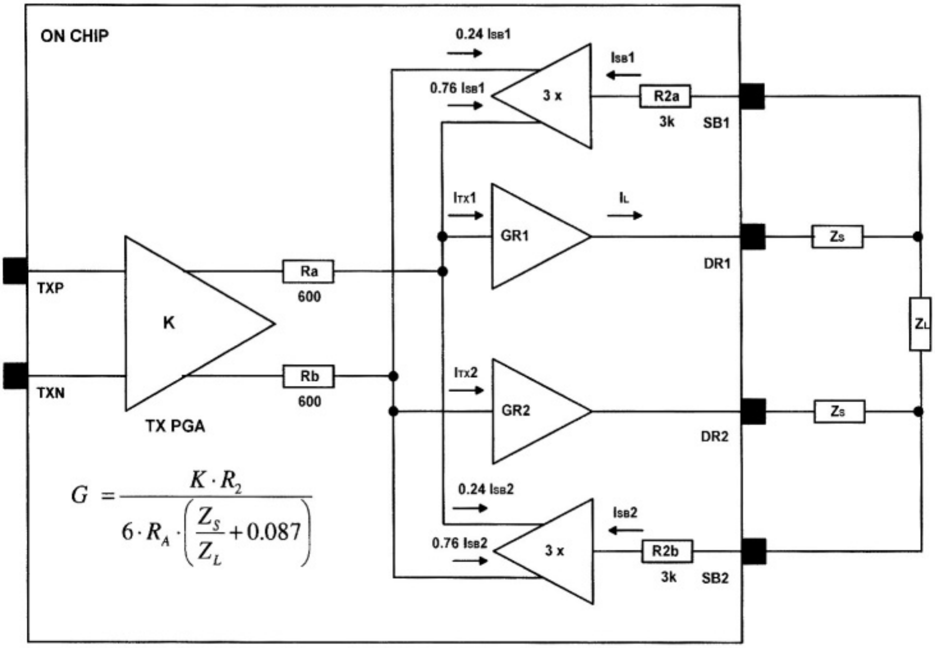

ON CHIP
0.24 ISB1
ISB1
0.76 ISB1
3 x
R2a
3k
SB1
ITX1
IL
GR1
Zs
DR1
TXP
Ra
600
K
ZL
TXN
Rb
600
ITX2
GR2
Zs
TX PGA
DR2
$G = \frac{K \cdot R_2}{6 \cdot R_A \cdot \left(\frac{Z_S}{Z_L} + 0.087 \right)}$
0.24 ISB2
ISB2
0.76 ISB2
3 x
R2b
3k
SB2

Fig.9. ADSL driver with current mode output stage block diagram

The driver circuit uses a low ohmic sense impedance (Zs) in series with the output. The voltages on each side of this impedances are sensed, converted to current and fed back to the inputs of current amplifiers GR1 and GR2 which transforms the sense impedances to the appropriate termination impedance for an ADSL application. Beside that, this feedback also determines the closed-loop current gains of the current amplifiers. With the synthesis factor m = 5.7 and equation (4) the differential output impedance can be calculated.

$$Z_O = 2 \cdot m \cdot Z_S \,. \tag{4}$$

Due to the programmable gain K of the preamplifier, the overall transmit gain can be varied between 12,9dB and 16,4dB in steps of 0,5dB. The output voltage of the pre-amplifier is converted to current via resistors Ra, Rb and fed into the virtual-ground input nodes of output amplifiers GR1 and GR2. To achieve high gain and good noise performance, the preamplifier circuit is based on a folded-

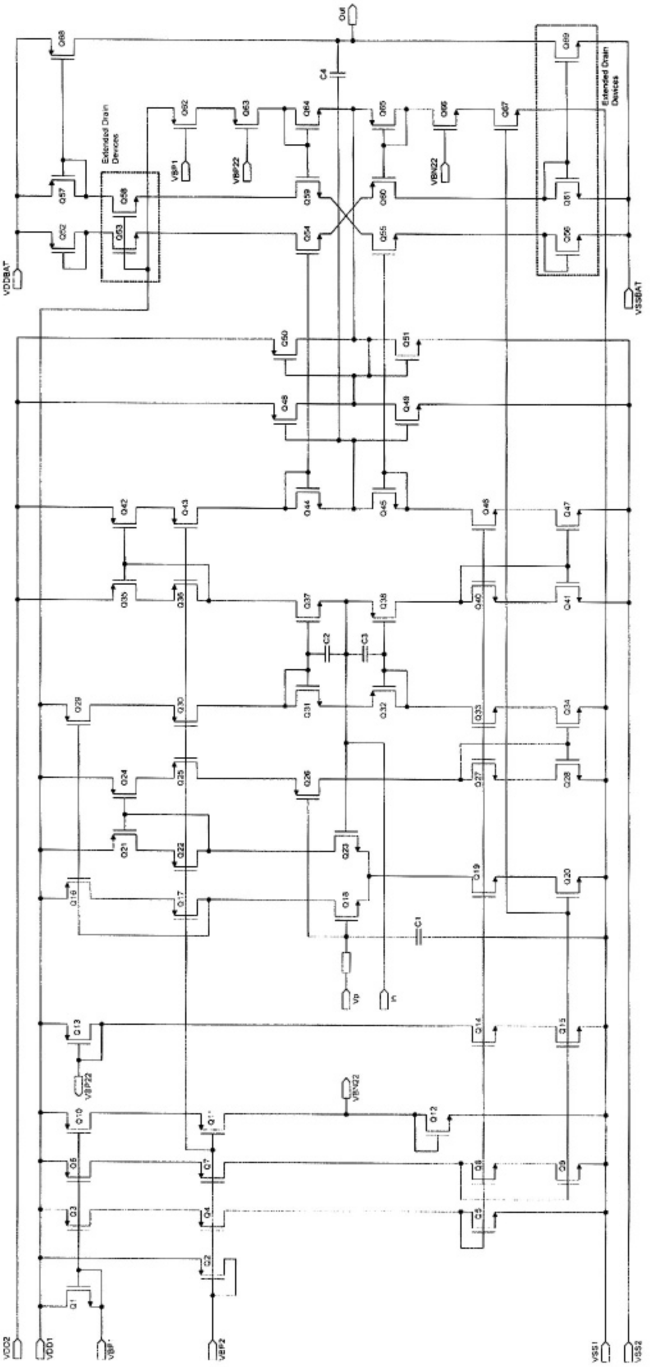

VDD2
VDD1
VBP1
VBP2
VSS1
VSS2
VDDBAT
VSSBAT
Out
Vp
Vn
VBP22
VBN22
VBP1
C1
C2
C3
C4
Extended Drain Devices
Extended Drain Devices

Fig.10. Schematic diagram of current mode buffer

cascode opamp structure. The line driver device requires two power supplies: Up to 11V for the output stage, and 5V for the remaining blocks.

Fig. 10 shows the simplified schematic of the used current mode output amplifier. To ensure a linear voltage to current conversion via resistors Ra and Rb, the input impedance has to be low. Therefore an operational transimpedance amplifier (OTA) with input stage Q18, Q23 followed by output stage Q37, Q38 has been implemented. The input signal current that appears on the sources of Q37 and Q38 passes through their drains and current mirrors Q41-Q47, Q35-Q42 and is then fed into the high impedance node. The resulting signal voltage is applied to the gates of transconductance-stages Q54, Q55 and, inverted by the action of ‘analog inverter’ Q48- Q51, to gm-stages Q59 and Q60, respectively. This arrangement is known as ‘Fischer-Koch’ structure. The drains of Q59 and Q60 are the current outputs of this structure. They are connected to the final current mirrors composed of transistors Q57, Q68 and Q61, Q69. A current transfer ratio of 1:50 at these mirrors ensures adequate output current drive. To withstand 11V supply voltage at the final output mirrors, extended drain devices have been used for transistors Q53, Q56, Q58, Q61 and Q69. Cascode transistors Q53 and Q58 protect the 5V output devices of the ‘Fischer-Koch’ structure. Capacitor C4 along with the output impedance at the high impedance node formed by the drains of Q43 and Q46 determine the dominant open-loop pole. The total quiescent current drain for this current mode output amplifier is 18mA. It is capable of driving 300mA into a 25Ω differential load.

5. A line-driving buffer used in a broadband HV-SLIC [4]

This line-driving buffer is part of a monolithic Broadband Subscriber Line Interface Circuit (B-SLIC), which combines the functionality of an analog 150-V ringing SLIC with that of a line driver for ADSL-Lite data signals. It is fabricated in the Smart Power Technology SPT170B, that offers DMOS and HV-PMOS devices for supply voltages up to 170V.

The following main requirements are derived from ADSL-Lite and POTS standards and from system simulations:

- Output voltage range: 60 to 70V for data and/or voice, up to 140V in ringing mode
- Output current up to 150mA
- Total Harmonic Distortion below -60dB for peak signals (25Vpp single ended), below -70dB for effective signal voltages up to 550kHz
- Slew Rate 100V/□s
- Noise Spectral Density: < -130dBm/Hz @ 100□
- Stability in all modes with all possible line impedances

- Robustness against over voltages
- Effective thermal protection

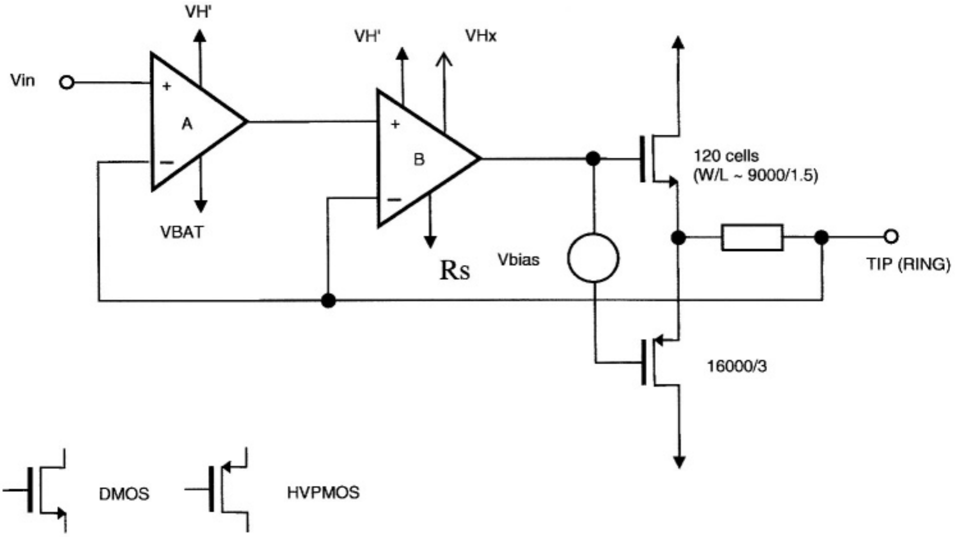


Fig.11. Buffer structure

The buffer structure finally chosen is that of Fig. 11. Because of the voltage and current demands, the output transistors have to be large DMOS or HV-PMOS devices. The output stage with the best power/linearity tradeoff then is the complementary source follower circuit in class-AB operation. Nevertheless, high gain of the driving circuit is still necessary over the whole signal frequency range, and that's why a nested loop structure with op-amps A and B is used. The goal was an overall bandwidth of about 20 MHz. But to achieve stable operation, the inner loop has to be significantly faster. However, small signal considerations are not sufficient. Slew rates have to fulfill similar conditions to avoid typical large signal problems like unstable step response.

High bandwidth together with high load capacitances calls for large transconductances, and so bipolar npn transistors have to be used in the signal path of opamp B (Fig. 12). They build the input differential pair Q1, Q2, while emitter followers Q3 and Q4 drive the large output transistors Ml and M2, respectively. In normal voice or data operation their collectors are connected to GND. They must not be switched to the positive supply VH, if the output voltage is low, because then the voltage across the transistors could exceed the allowed limit (90V). On the other hand, VH is required to enhance the output

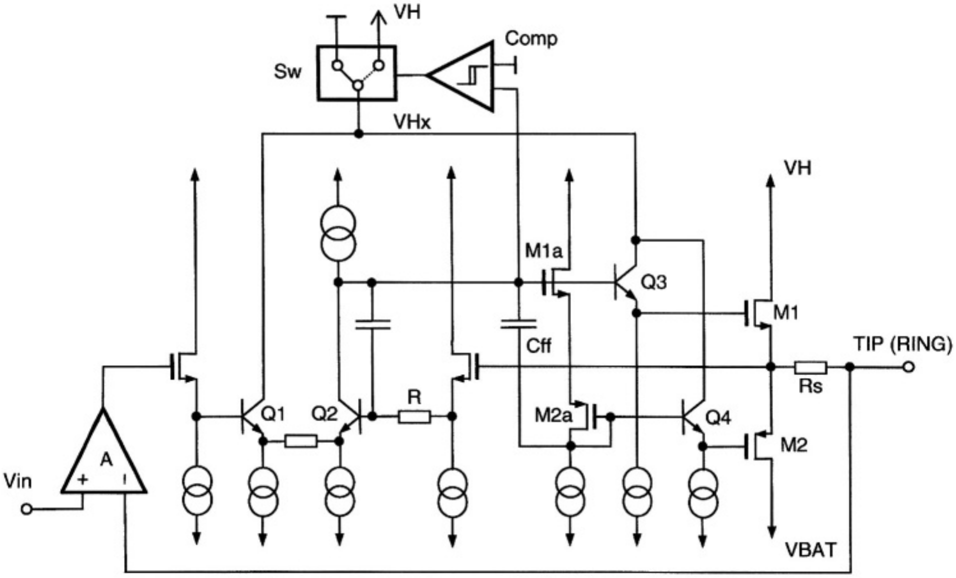

VH
Sw
Comp
VHx
VH
M1a
Q3
M1
Cff
TIP (RING)
Rs
R
Q1
Q2
M2a
Q4
M2
A
+
Vin
VBAT

Fig.12. Buffer with details of opamp B

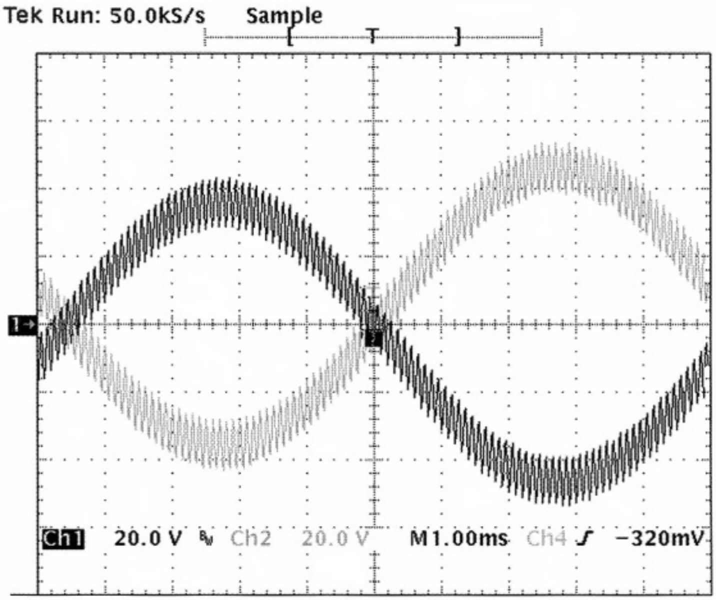


Fig.13. Measured 50kHz/20Vpp sine wave superimposed on 60Vrms/100Hz "ring" signal

voltage range in the ringing mode. For output voltages above GND, the collector potentials have to be switched to VH, and they have to be switched back with the negative zero crossing in order not to endanger the npn transistors. Therefore the switch SW together with comparator Comp has been introduced. The additional distortions arising from this switching could be kept short and small enough as not to disrupt data transmission during ringing. Fig. 13, measured with a large 60Vrms, 100Hz "ring" signal with 20V DC offset, that is superimposed by a 20Vpp, 50kHz "data" sinewave, is intended to illustrate this situation.

Stabilizing the buffer turned out to be a key problem. Extensive circuit simulations preceded the final solution. Somewhat simplified in Fig. 12, R1 and C1 are intended to provide pole-zero cancellation, while Cff forms a feedforward path across the high voltage MOS diodes M1 and M2, that as usual define the quiescent current of the output stage.

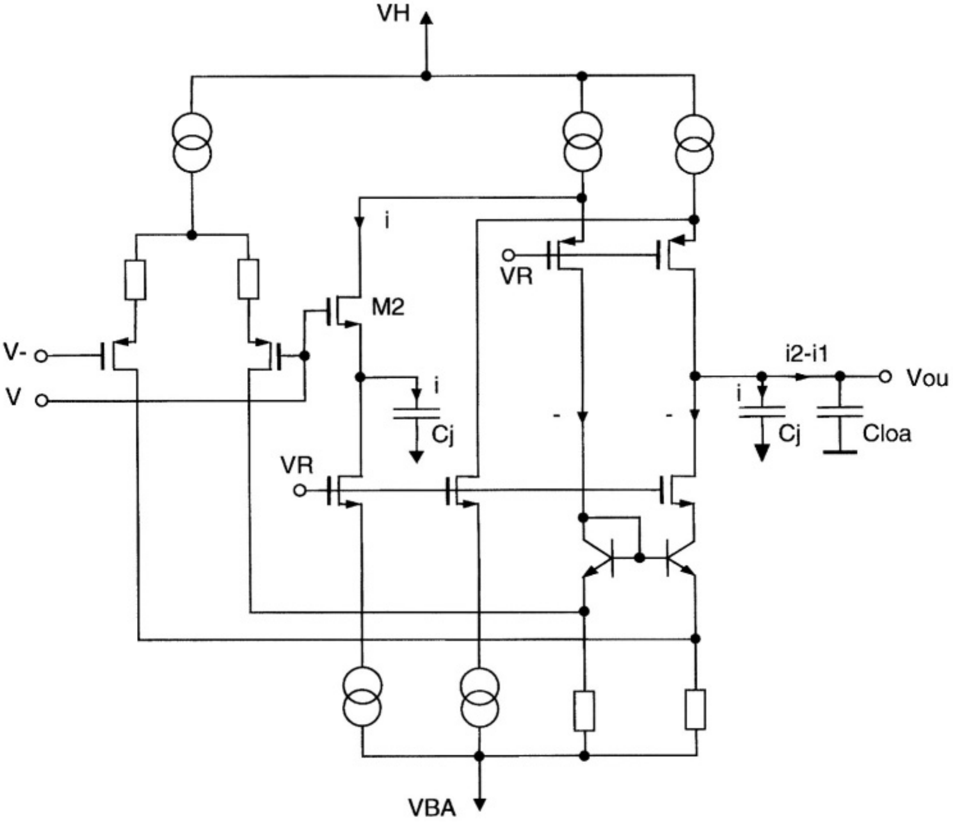


Fig.14. details of opamp A

Opamp A is a “double folded” BiCMOS circuit with HVPMOS input pair (Fig. 14). The most interesting aspect of this structure probably is a technique to compensate for critical junction capacitances with their nonlinear voltage dependence. Here, the current flowing into the load capacitance Cj1 at the output node can be compensated by adding a similar capacitance Cj2 and transistor M2. Due to M2 and the unity gain configuration, the voltage at the compensation capacitance nearly equals the output voltage. The corresponding capacitive current is subtracted at the output node, thus in first order cancelling the effect of Cj1. From simulation results, this measure yields a 10dB improvement in the opamp’s distortions to -80dB values.

6. Class AB driver efficiency

Driver efficiency is defined as the output power Po delivered to the line divided by the power dissipated in the driver PD plus the output power. For an ideal driver with the maximum peak output voltage equal to the supply voltage and zero quiescent current, the efficiency can be easily calculated, if the output voltage Vo and the supply voltage Vs are known.

$$P_D = \frac{1}{T}\int_0^T \frac{(Vs - v_o(t))\cdot |v_o(t)|}{R_L}\,dt \tag{5}$$

$$P_O = \frac{{V_{Orms}}^2}{R_L} \tag{6}$$

$$\varepsilon = \frac{P_O\cdot 100\%}{P_D + P_O} \tag{7}$$

From calculating the integral in (5), equation (7) can be simplified to

$$\varepsilon = \frac{{V_{Orms}}^2\cdot 100\%}{V_S\cdot \overline{v_o(t)}}, \tag{8}$$

where $\overline{v_o(t)}$ is the mean absolute value of the output voltage. Knowing the amplitude X of a DMT signal to be Gaussian distributed, with mean value 0 and standard deviation σ equal to the effective signal value X_{rms}, we can substitute the time integration by an integration over the probability weighted signal amplitudes [4].

$$\overline{v_o(t)} = \int_{-\infty}^{+\infty} |x| \cdot p(x) \cdot dx = \frac{1}{\sigma \cdot \sqrt{2 \cdot \pi}} \int_{-\infty}^{+\infty} |x| \cdot \exp\left(-\frac{(x-\mu_0)^2}{2 \cdot \sigma^2} \right) \cdot dx \tag{9}$$

For a pure ADSL signal without dc current flow, $\mu_0 = 0$, and the evaluation of the integral yields

$$\overline{v_o(t)} = \sqrt{\frac{2}{\pi}} \cdot \sigma . \tag{10}$$

Substituting $\overline{v_o(t)}$ in equation (8) and taking into account that the supply voltage Vs must be at least equal to the crest factor times the RMS value of the output voltage Vo, equation (8) yields

$$\varepsilon(DMT) = \frac{{V_{Orms}}^2 \cdot 100\%}{V_{Orms} \cdot CF \cdot V_{Orms} \cdot \sqrt{\frac{2}{\pi}}} = \sqrt{\frac{\pi}{2}} \cdot \frac{1}{CF} \cdot 100\% . \tag{11}$$

In a realistic application we have to consider at least an active termination of the line with a certain synthesis factor m and a drop voltage V_D in the output stage of the driver, which increases the necessary minimum supply voltage. In addition, the driver circuit itself will consume a certain amount of quiescent current I_{QR}. Due to this we have to expand equation (11) to

$$\varepsilon(DMT) = \frac{100\%}{CF \cdot \sqrt{\frac{2}{\pi} + \frac{V_S \cdot I_{QR}}{P_O}}} \cdot \frac{m}{m+1} \cdot \frac{V_S - V_D}{V_S} . \tag{12}$$

To give an example, the efficiency of the driver circuit described in section 3 is calculated. With $CF = 4.2$, $V_S = 15V$, $I_{QR} = 2.5mA$, $m = 6$, $V_D \approx 3V$ and $P_O = 19.8dBm = 95.5mW$,

$$\varepsilon(DMT) = \frac{100\%}{4.2 \cdot \sqrt{\frac{2}{\pi} + \frac{15 \cdot 0.0025}{0.0955}}} \cdot \frac{6}{6+1} \cdot \frac{15-3}{15} = 18.32\% . \tag{13}$$

7. Conclusion

Three different class-AB driver circuits for CO ADSL applications, fabricated in complementary bipolar, CMOS, and HV-BiCMOS processes have been presented. The efficiency of a class-AB line driver suffers from the high crest factor appearing in DMT modulated DSL systems. Due to the lower bill of material for the required external components and the smaller chip area compared to a class-G line driver, class-AB line drivers are still a competitive option, especially in DSL systems with crest factors lower than 4,5.

References

[1] D. Smith, M. Koen and A. F. Witulski, "Evolution of high-speed operational amplifier architectures," *IEEE J. Solid-State Circuits,* vol. 29, no. 10, Oct. 1994

[2] S. Franco, "Current-feedback amplifiers benefit high-speed designs", *EDN,* pp. 161-172, Jan. 5, 1989.

[3] D. F. Bowers, "The impact of new architectures on the ubiquitous operational amplifier", in *Proc. of the 1992 Workshop on Advances in Analog Circuit Design in Europe.*

[4] B. Zojer, R. Koban, J. Pichler and G. Paoli, "A Broadband High-Voltage SLIC for a Splitter- and Transformerless Combined ADSL-Lite/POTS Linecard", ," *IEEE J. Solid-State Circuits,* vol. 35, no. 12, Dec. 2000

Class D self-oscillating Line Drivers

Tim Piessens, Michiel Steyaert
KU. Leuven, ESAT-MICAS
Kasteelpark Arenberg 10
3000 Leuven, Belgium

Abstract

Power efficiency is one of the most important specifications for the design of an integrated line driver. Since the efficiency of a linear class AB line driver is inverse proportional with the crest factor of the signal, these type of line drivers have low efficiencies for signals with non-constant envelope modulation. These crest factors are not only be found in low frequent audio applications, but also in high data-rate communication systems like ADSL and VDSL modems. Switching type line drivers have the basic advantage that next to a parasitic on-resistance of the switches no power is dissipated in the active elements. So theoretically an 100% efficiency independent on the crest factor can be obtained. Two types of class D amplifiers are considered: synchronised and self-oscillating drivers. Originally self-oscillating class D power amplifiers were designed to reduce the number of components but recent research revealed important properties of this asynchronous modulation which increases the systems efficiency, linearity and bandwidth.

1. Introduction

The design of high efficiency power amplifiers is an old problem. Due to the stringent linearity and bandwidth specifications, most common systems start with a linear type class AB power amplifier. However its efficiency for a signal with a Gaussian amplitude distribution is given by :

$$\epsilon = \frac{V_{DD}^2\sqrt{\pi}}{CF\ V_{DD}(V_{DD}\sqrt{2} + CF\ I_q\sqrt{\pi}) + CF\ P_{qcc}\sqrt{\pi}} \tag{1}$$

In which V_{DD} denotes the supply voltage, I_q the quiescent current and P_{qcc} the power dissipated in the quiescent current control circuitry. The most important factor in this equation is the Crest Factor (CF), being the ratio of the peak voltage divided by the rms voltage of the applied signal.

A. van Roermund et al. (eds.), Analog Circuit Design, 309-331.
© 2003 *Kluwer Academic Publishers. Printed in the Netherlands.*

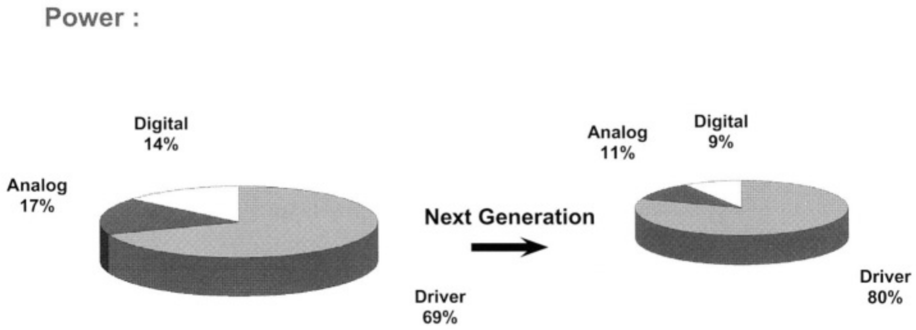


Figure 1: Relative portion of the power consumption budget for different building blocks of a commercial Central Office (CO) chip-set and its evolution towards a next generation.

For a Gaussian signal, the minimum CF is 2.5, so the maximum efficiency is 50%, and quickly degrading for real signals. For instance, audio signals like real music fragments, have CFs between 10 dB and 20 dB with a strong average around 15 dB [1]. Taking a look to novel modulation schemes like Discrete Multi-Tone modulation (DMT) and Orthogonal Frequency-Division Multiplexing (OFDM), the CFs associated with these modulation are again above 15 dB. These modulation techniques were developed in order to fully deploy lossy channels like the twisted pair copper wires used in the xDSL application fields [2–4]. In this paper mainly the line driver problems in Digital Subscriber Loop (xDSL) applications are addressed since this is the present driving application in this research field, but keep in mind that many of these ideas originate or arc also applicable in the audio field.

To indicate the importance of line driver efficiency for an xDSL system, figure 1 shows the comparison between the relative power consumption of two typically Asymmetrical Digital Subscriber Loop chip-sets. For this the values from different commercial chip factories were taken and a relative mean has been calculated. To the right of figure 1 the same calculations were redone but then for the consecutive chip-sets of the same foundries. Surprisingly, the total power budget does not seem to decrease significantly from one generation to the next. A closer inspection reveals the line driver as being the real bottleneck. In the most recent generation of xDSL chip-sets the relative portion of the line drivers power consumption has increased to 80% of the total power budget.

In the next section the different considerations for building high efficiency, switching line drivers for xDSL are further elaborated. The basic properties of

Table 1: Summary of the most important xDSL requirements for line driver design. The V.34 standard is added in grey for comparison reasons.

	Distortion	*Bandwidth*	*Output Power*
V.34	THD <-70 dB	4.96 kHz	0 dBm
ADSL(-Lite) US‡		103.5 kHz	13 dBm
ADSL-Lite DS†	MTPR > 34 dB	418.3 kHz	16.3 dBm
ADSL DS†	MTPR > 55 dB	970.3 kHz	20 dBm
VDSL DS†	MBD > 63 dB	8.5 MHz	14.5 dBm

† DS = downstream, ‡ US = upstream

a DMT modulated signal and the requirements for a working xDSL system will be combined to derive some basic limits for the design of an xDSL line driver. It will be made clear that in order to reach a sub 500 mW consuming line driver, switching type amplifiers should be investigated.

In section 3., different synchronised class D power amplifiers are further explored. The Self-Oscillating Power Amplifier (SOPA) will be the subject of section 4..

2. Wire-line Communication

2.1. xDSL-Technologies

In this section we will mainly focus on the ADSL and VDSL part of the xDSL family. Not only for their dominant market position, but also since they depend on DMT-modulation techniques and are thus the most demanding technologies for the line drivers [5].

In table 1 the most important properties for line driver design of ADSL and VDSL are summarized. To compare these specifications with older technologies, one can clearly notice the enormous shift in line driver requirements for the two digital subscriber modem types. In less than one decade of xDSL research activities, an improvement of over 3 decades in bit-rate has been reached. This however at a drastic cost in the Analog Front-End (AFE) design specification. The linearity specifications are not significantly relaxed, but the bandwidth and necessary output power are increased towards the current technology limits.

The requirement for a minimal power dissipation is a direct consequence of the deployment scenarios of xDSL communication. In an xDSL system, the Customer

Premises Equipment (CPE, = equipment at the users side) is with a single twisted copper wire-pair connected to the Central Office (CO). Since the attenuation of a copper wire is the limiting factor for the achievable bit-rate, the loop-lengths have to be kept short. For customers who reside too far from the telephony central, the CO-modem has to be deployed from a street-side cabinet case. This is mostly the case for VDSL-modems since loop lengths are becoming very small, to reach the high bit rates for this technology. These cabinet cases don't have temperature control except for natural heat convection which is limited due to its small volume. The maximum operating temperature and thus maximum heat dissipation limits the amount of install-able lines and thus the number of servable customers.

The stringent linearity requirements are a direct consequence from the fact that the twisted copper wires are not shielded. Therefor the out-of-band specifications of the xDSL modem are very stringent to avoid cross-talk to other twisted pairs in the bundle and e-gress radiation into other frequency bands like the ones used for amateur radio.

For the choice of a process technology, one has to take into account that the twisted pair is also shared with the Plain Old Telephony System (POTS). On the copper wire high voltages occur due to the POTS system. A line interface that is able to handle POTS and xDSL on the same die, has to take these high voltages into account [6–8].

As a summary, due to the high performance of an xDSL system, the line driver has to:

- Drive a relative large power to the line
- with a high linearity
- in a high bandwidth.
- The power dissipation needs to be minimal to allow maximal install-able lines at the CO-side
- Out-of-band specifications are very stringent (<-100 dBm/Hz)
- Deal with high voltages

For its high linearity, the class AB line driver was the most popular xDSL line driver in recent chip-sets [7, 9–11]. However due to its low efficiency (< 15%), the number of lines per $500\ \mathrm{cm}^2$ board is limited to 24 in order to stay NEBS compliant [12]. A next step in the evolution of xDSL chip-sets is the introduction of multiple supply class G line drivers [13]. For a two supply class G power

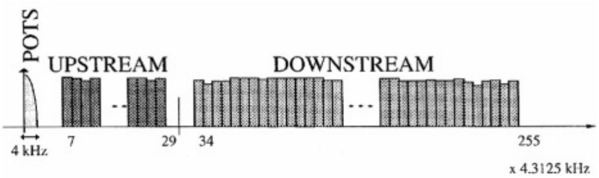


Figure 2: Spectrum of an ADSL system.

amplifier, the efficiency for driving a Gaussian signal is given by :

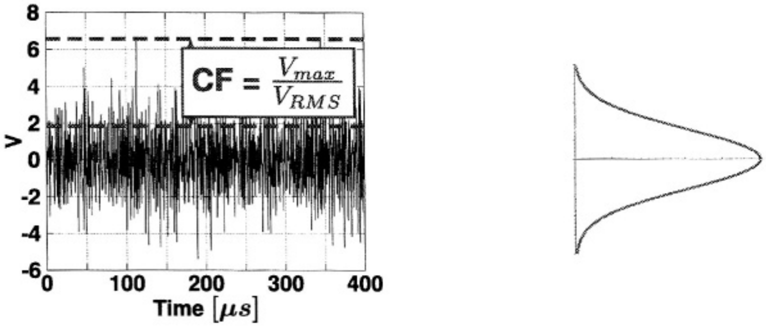
$$\epsilon = \frac{n\sqrt{\pi}}{CF\sqrt{2}\left(\left(1-\exp\left(\frac{-CF^2}{2n^2}\right)\right)+n\exp\left(\frac{-CF^2}{2n^2}\right)\right)} \qquad (2)$$

with n denoting the ratio between the high and low supply voltage. The efficiency shows a relaxed decay with the CF of the signal, since an optimal segmentation of the supplies will lead to almost 50% efficiency for the major part of the ADSL symbols. In this way, this solution relaxes the dissipation limit towards 48 lines per $500\ \mathrm{cm}^2$ board. This limit however, can only be broken by going to switching type line drivers.

2.2. Basic Properties of Multi-Tone Channels

Figure 2 shows a rough sketch of an ADSL spectrum. The Discrete Multi-Tone (DMT) modulation consists of 22 carriers containing the upstream (=from CPE to CO) information and 221 downstream (=from CO to CPE) carriers. Each carrier is Quadrature Amplitude Modulated (QAM). The constellation size of each carrier is given by the signal-to-noise ratio (SNR) of the channel at the specific carrier frequency [14]. Since the carriers are individually adaptable to different line conditions, it can be understood that this modulation scheme allows bit rates near the Shannon limit of the twisted pair information channel.

The time domain representation of the signal is given in figure 3 (a), together with its amplitude distribution. Since the constellation points can be regarded as random variables for normal data communication and the number of carriers is sufficiently large for the central limit theorem to hold, the amplitude distribution f_A of the DMT-signal $x(t)$ can be approximated as being Gaussian. From this assumption, the distribution function for the CF can be calculated. By calculating the extreme value distribution for a DMT generated symbol with N carriers, the



(a) (b)

Figure 3: Time domain representation of a DMT-modulated signal (a) and amplitude distribution (b).

distribution of the CF is obtained :

$$f_{CF} = \frac{2\sqrt{2}}{\sigma\sqrt{\pi}} N \left(\mathbf{erf} \left(\frac{\sqrt{2}CF}{2\sigma} \right) \right)^{2N-1} \exp\left(-\frac{CF^2}{2\sigma^2} \right) \tag{3}$$

The result of this calculation is plotted in figure 4 and compared with the histogram obtained from a monte carlo simulation. To obtain this histogram, the crest factor of 50000 randomly generated 4QAM modulated ADSL downstream signals has been calculated, showing a big accordance with the obtained model. This Fisher-Tippet distribution has a slow tail when going to high crest factors. The consequence of this slow tail is that a line driver needs to be designed so it can cope with high CFs, in order not to degrade the bit error rate too much.

In practical designs the crest factor is limited. The clipping of the CF will lower the quantisation noise of the used Analog-to-Digital Converter (ADC) and the Digital-to-Analog Converter (DAC) and their power consumption, but will introduce clipping noise [15]. In practical ADSL systems the CF is limited to a factor of 15 dB or 5.6.

3. Switching type line drivers (Class D)

3.1. Basic Class D configuration

A typical class D consists of three major parts: a Pulse Width Modulator (PWM) or Pulse Density Modulator (PDM) which transforms the analog input signal into

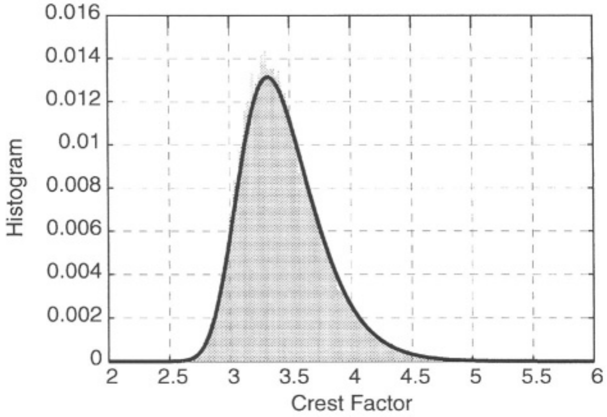


Figure 4: Distribution function for the crest factor of an upstream ADSL symbol (solid line) compared with a monte carlo simulation (grey histogram).

a switching signal, large output switches to deliver the current to the load and an output filter to get rid of the high frequent switching signals. A principle schematic is shown in figure 5. The distinction between synchronised and self-oscillating class D amplifiers is made whether the modulator is clocked or not. In this section only the first type of class D power amplifiers is considered.

The major advantage of the class D principle is the use of a switching output stage. If the on resistance of the output switches are neglected, there is no simultaneous current through the active element and voltage standing over the active element. So no heat is dissipated in the switches, opening the opportunity for 100% efficiency. For its high efficiency the class D power amplifier is very important in the low power consumption audio application field like hearing aids, etc.. To use one as a line driver several extra considerations need to be taken into account:

- For audio signals the output filter does not need to be very steep for audio signals have a low bandwidth (<20 kHz) compared with the switching frequency (usually between 200 kHz and 500 kHz). Moreover the loudspeaker set (and eventually also the human ear) inherently is a low pass filter for those switching frequencies.
- xDSL line drivers have relative high bandwidths and very stringent out of band specifications. Therefor the output filter needs to be very steep or the over-switching ratio (OSR = mean switching frequency / signal bandwidth)

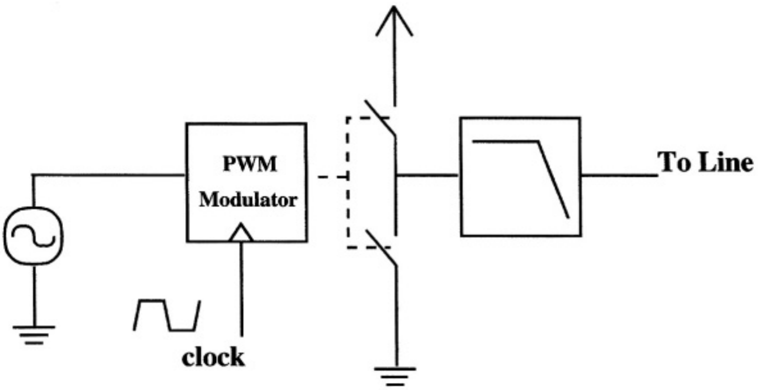


Figure 5: Principle schematic of a class D power amplifier

has to be set high enough. The output filter needs to be a passive filter, so for line driver applications it will require many extra passive components, lowering the integrate-ability and increasing the cost of the complete driver.

- Another consequence of a higher switching frequency is that the switching losses will become more important in the overall power consumption. So much care needs to be taken to optimally choose the switching frequency in order to still gain efficiency.

From this brief analysis it is to be concluded that the use of a technology which already matured in the audio field as a line driver opens up a whole set of new design consideration.

3.2. Output Stage Considerations

On-resistance

The design of the output stage is a far from trivial task. From a power point-of-view several contributions need to weighed against each other. The first important parameter is the conduction loss due to the non-zero on-resistance of a real switch. When integrating these switches, this will lead to bigger transistors. An alternative would be the use of a higher supply voltage. The load resistance can be transformed to higher values for the same output power. This problem becomes more pronounced when integrating a line driver in a low voltage technology.

Non-zero switching time

Another important efficiency degrading effect is the relative importance of the switching losses in the overall power budget. During a switching event, there is a current flowing through the switch, while a voltage over the switch is building up. This leads to extra power dissipation. For this the switching time needs to be as low as possible. This can be obtained by reducing the output transistor sizes, which is the opposite action from the previous degradation effect. Another approach would be reducing the mean switching frequency, in order to reduce the relative portion of the switching losses. For the class D to be used as a line driver, the large bandwidth could be problem. If the on-resistance of the switch is small, an almost short circuit is made between supply and ground during the transition, giving an extra power consumption. This shoot-through current can be cut by inserting a dead time between the switching events of both switches. This however leads to an important in-band cross-over distortion, since the output voltage is uncontrolled during the dead zone.

Dynamic power consumption

Closely connected with the previous observations is the dynamic power consumption. Huge switches tend to have large input capacitances which need to be charged and discharged at every switching event. Furthermore, in order to reduce the switching time, these large input capacitances need to be driven by buffer amplifiers whose power consumption will increase with decreasing output switching time and larger input capacitance of the final stage.

In-band signal integrity

The class D configuration of figure 5, shows an open loop class D power amplifier. Due to the switching nature of the output signal, feedback is not that trivial. If no feedback is foreseen, the in-band signal integrity is dependent on the power line integrity. Every spur on the supply rail will be directly transferred towards the output. Since power rail bounce is directly related by the power consumption, a regulator with enough decoupling should be added on the supply rail which has an equally strict linearity as the required output linearity.

Another source of distortion in an open loop class D line driver is the cross-over distortion which occurs during an output transition. The cross-over distortion does not only originate from a no shoot-through current circuitry, but also from a the inductance of the output filter. Since the current through an inductor can not stop immediately, the remanent inductive load current flows through a free-running diode. The diode voltage adds to the supply, so on the output a voltage occurs which is smaller than ground or bigger than the upper supply rail. These effects

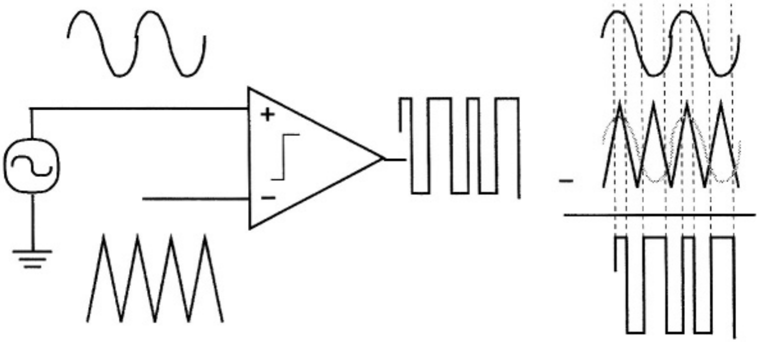


Figure 6: Principle schematic of a basic open loop modulator block.

become more pronounced for higher switching frequencies, for their relative importance on the integrated output voltage (through the output filter) will increase.

Other Remarks

Other important specification for the design of a class D output stage are of more process technological nature. The diodes need to have fast recovery times and the technology should have a high latch-up resistance when these diodes conduct the remanent inductive current.

As it has already been pointed out, the use of higher output voltages is always beneficial for the design of a class D line driver. This can be accomplished by using a more expensive DMOS technology [8], which reduces the integrate-ability of the line driver with other system building blocks. Another approach could be the use of drain-source engineering in standard CMOS technologies [16]. When a high voltage is not feasible, high currents will occur in the output path. Careful designs is mandatory in order to avoid ‘hot spots’ and electro-migration of the metal. Techniques from clock distribution networks need to be used in the layout of a segmented output switch to guarantee an equal current spread.

3.3. Modulation Schemes

Figure 6 shows the basic and most common open loop modulation technique, which is called natural sampling. It was used for its simplicity and for the absence of any extra digital processing. Normally, no harmonic distortion is generated by this technique. The modulation however generates spurs at multiple harmonic offsets from the central clocking frequency. This limits the bandwidth of the system.

Another big disadvantage is the absence of a feedback loop, so all circuit non-idealities like output stage inaccuracies, non-linearity, timing errors or supply voltage ripple cannot be compensated. There exist techniques to allow the insertion of feedback to the natural sampling technique, but:

- The output needs to be filtered before it is fed back. The slew rate of this filtered output needs to be much smaller than the slew rate of the triangular wave. Otherwise extra higher harmonics will be generated driving the amplifier in a high frequent self-oscillating mode. The mean switching frequency is pushed to higher frequencies with all associated problems for the output stage.
- As a consequence of this low corner frequency of the feedback filter is a low loop gain and thus a low compensation of output errors. This limits the performance improvement.
- The output filter cannot be made sharp enough to filter out the switching component completely, since the parasitic resistance limit the filter order. This remaining switching component will be amplified by the pre-amplifier and will arrive at the comparator were it most certainly will be out-phased with the original triangular wave. The intermixing products will generate in-band distortion.
- High frequent ripple picked up from the environment at the inputs of the comparator can cause erroneous switching of the class D amplifier which cannot be corrected by feedback, since it is filtered out in the correcting loop. The comparator and pre-amplifier therefor have to be shielded very carefully.
- Pulse amplitude errors are due to changes in the output load. These changes are very common in wire-line communications.

Another approach would be the use of digital modulation techniques like **$\Delta\Sigma$-modulation** to create a PWM signal [17]. In this technique the errors are shaped towards higher frequencies. An external filter needs to filter out these high frequency errors. The major advantages of these techniques are

- The accuracy of the analog components is largely relaxed, compared with the PWM modulator.
- Since the high frequency errors are randomised, the $\Delta\Sigma$ modulation generates less in-band substrate noise and thus less self-interference.
- Baseband feedback is inherent in the system, so in-band distortion is lower for the same clocking frequencies.

- For low signal bandwidths this can be combined with an Analog to Digital Converter (ADC) if the base signal is already digital [18].

Formula 4 calculates the Dynamic Range (DR) of a $\Delta\Sigma$ based switching power amplifier.

$$DR = \frac{3\pi}{2}(2n+1)\left(\frac{OSR}{\pi}\right)^{2n-1} \tag{4}$$

The $\Delta\Sigma$-loop has order n and an over-switching ratio (OSR). The order should be kept as low as possible, since the order of the output filter needs to be at least one order higher than the $\Delta\Sigma$ order. So in order to meet the necessary linearity specifications, a higher OSR has to be chosen. This however, requires very high frequent high-power devices, which are not available in present semi-conductor technologies.

4. Self Oscillating Class D line drivers

4.1. Basic SOPA architecture

The construction of a self-oscillating power amplifier (SOPA) originates from the observations of the natural sampling scheme of figure 6. The easiest way to generate a triangular wave is by designing a relaxation oscillator. Since the building blocks of a relaxation oscillator are the same as the ones for the natural sampling modulation technique, a merge of the two circuits seems a logical step. The result of this is shown in figure 7. In the same figure the typical waveforms are plotted. They can be interpreted as follows: since the filtered version of a square wave is a triangular wave, a triangular wave is most likely to be found at the negative input of the comparator. Due to the feedback, the input signal can be considered as a shift in bias point, so a triangular approximation of the input signal resides at the negative input node. Since for in-band signals the system is in unity feedback and the loop filter is assumed to be linear, the in-band frequency content of the output square wave will be the same as that of the triangular wave and thus as the input-signal.

Although the circuit was firstly conceived to save components in the audio field, in the next other more important properties will be shown in the next of this paper. Important properties of this circuit are that no clocking is used, so the loop can be considered to be fully analog and that inherent feedback is present to compensate output non-idealities. The mean switching is governed by the limit cycle frequency, i.e. the frequency of the oscillation that occurs when no input signal is applied. Due to the filtering of the higher frequencies by the loop filter, the

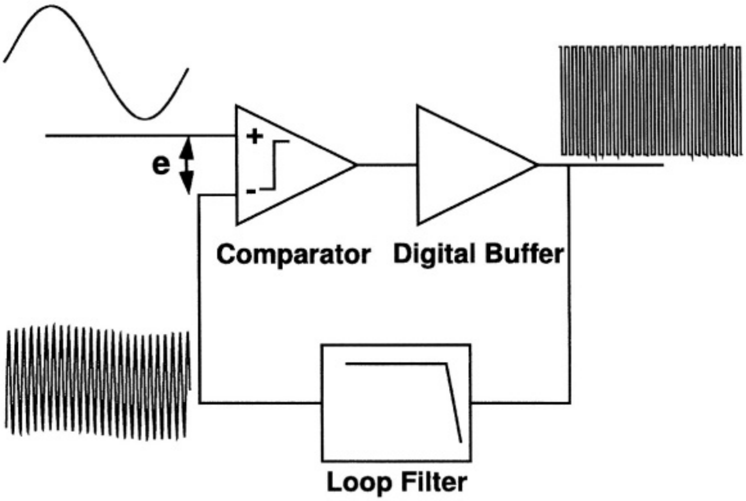


Figure 7: Principle schematic of a 0^{th} *order SOPA with characteristic waveforms.*

describing function analysis technique [19] can be used to analyse the hard non-linearity. The comparator is modelled in the frequency domain as the following single input describing function :

$$N(A) = \frac{2V_{DD}}{\pi A} \tag{5}$$

The term A represents the limit cycles amplitude at the input of the comparator. The Barkhausen criterion for self oscillation with the loop filter *L(s)* is:

$$1 + N(A)L(s) = 0 \tag{6}$$

Solving this complex equation gives us an expression for the limit cycle amplitude and frequency. The transfer function of the linearised system for error signals with amplitude e is given by

$$T(s) = \frac{N_2(A,e)}{1 + N_2(A,e)L(s)} \tag{7}$$

In which the dual-input describing function $N_2(A, e)$ describes the frequency domain input-output relation of the comparator in presence of a limit cycle oscillation with amplitude A for the forced signal with a given frequency and an amplitude at the comparator input e. In the closed loop SOPA system e will be the error signal. The describing function $N_2(A, e)$ can be calculated as follows:

$$\begin{aligned} N_2(A,e) &= \begin{cases} \frac{V_{DD}}{\pi e}\left(\frac{e}{A}\right) {}_2\mathrm{F}_1\left(\frac{1}{2}, \frac{1}{2}; 2; \left(\frac{e}{A}\right)^2\right) & \text{for } 0 < e < A \\ \frac{2V_{DD}}{\pi e}\, {}_2\mathrm{F}_1\left(\frac{1}{2}, -\frac{1}{2}; 1; \left(\frac{A}{e}\right)^2\right) & \text{for } 0 < A < e \end{cases} \\ &\approx \quad N(A)/2 \qquad \text{when } 0 < e << A \end{aligned} \tag{8}$$

Note that the error signal e. denotes the amplitude measured at the input signal frequency, while A denotes the amplitude at the limit cycle frequency. The describing function technique splits the signals at the different frequencies under consideration. In this way different amplitudes can be regarded at the same spatial location. From equation 8 the following qualitative conclusions can be drawn:

- if the error signal is reduced by the feedback loop in such a way that the error signal amplitude e becomes smaller than the limit cycle amplitude A calculated from (6), the transfer function becomes independent on the error signal itself. Thus the switching comparator behaves as a linear amplifier for the input signal. Equation (8) in that way imposes also a limit on the maximum output signal swing of the system.

- since for low error signals e, the gain of the comparator reduces to the single sinusoid describing function $N(A)/2$. The loop transfer function thus always have a fixed point in the complex planes governed by the limit cycle amplitude and frequency. The limit cycle oscillation thus acts as a natural dither who measures the output inaccuracies by the Barkhausen criterion (6) and adapts the loop transfer function to counteract the change in output characteristic. This phenomenon is called self-adaptivity of a non-linear circuit [19] and is very useful in a line driver where the output impedance can vary due to the environment. This is advantageous compared with the natural sampling technique.

- Since the input signal is approximated by a continuous time, discrete amplitude signal and not by a discrete time, discrete amplitude signal as in the digital modulation schemes. This will lower the quantisation noise compared with e.g. $\Delta\Sigma$-modulation.

4.2. Coupled SOPA amplifiers

A complete SOPA line driver is a bridge-like connection of two basic SOPA amplifiers. The line transformer, which is always present in wire-line communications

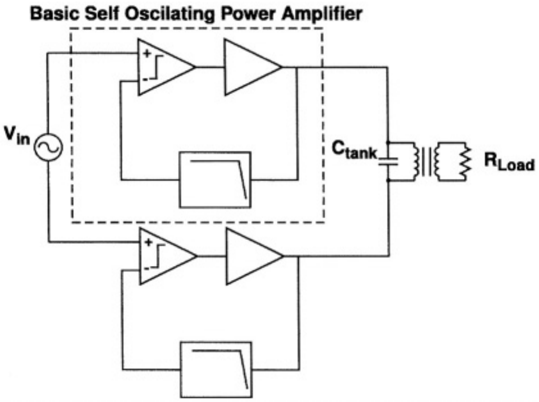
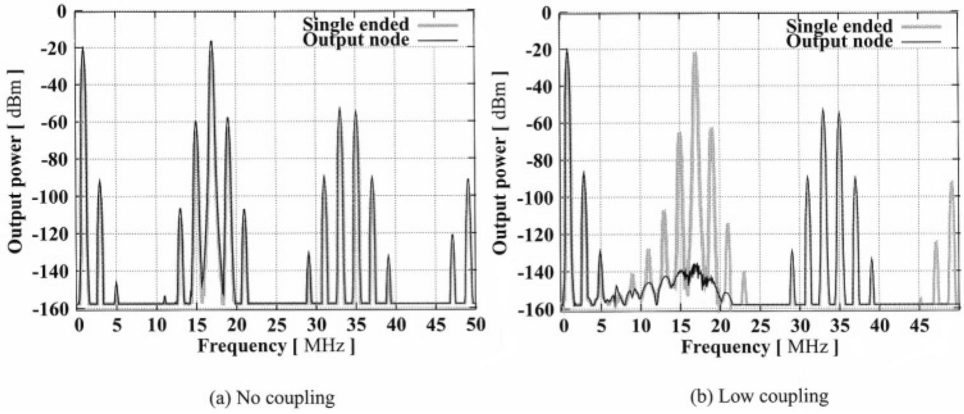


Figure 8: Full SOPA line driver coupled using a line transformer.

for galvanic isolation, forms together with the parasitic output impedance of the output stage, the coupling between the two basic SOPA amplifiers. The bridge connection allows a doubled output-swing for the same supply voltage. However, the major advantage of this configuration is the phenomenon of oscillator pulling between the limit cycles of both basic SOPA building blocks. Since both SOPAs are un-clocked oscillator pulling towards common mode is always the dominant oscillation mode even for low coupling factors. In this way the mean switching frequency becomes common mode for the primary of the line transformer and thus filtered out. The simulated spectra for an un-coupled SOPA and a coupled SOPA are given in figure 9. Since the limit cycle frequency becomes common mode due to the oscillator coupling, it is completely suppressed at the output. It can also be shown that the modulation peaks in the odd limit cycle bands are at even multiples of the input frequency [20]. Since the input signal is counter mode, these even harmonics will be common mode at the output. In figure 9(b), it can be clearly seen that the complete odd limit cycle bands are suppressed. This roughly doubles the signal bandwidth and relaxes the output filters specifications, since the energy in the even limit cycle bands is much lower than in the odd bands.

4.3. Higher order improvement

The previously described SOPA is often called a zeroth order line driver. Distortion levels of this structure are only dependent on the signal levels on the output



(a) No coupling

(b) Low coupling

Figure 9: Simulated spectra of an un-coupled (a) and coupled (b) SOPA line driver. Single ended spectra are plotted in grey.

node, and are thus only determined by the supply voltage. To improve the linearity of the line driver, a noise shaping technique, not unlike the one used in $\Delta\Sigma$-converters is introduced in the SOPA principle. Figure 10 shows a simplified model that will be used to explain the noise shaping technique. If we calculate the transfer from the input to the output, we will find :

$$\frac{V_{out}}{V_{in}} = \frac{H(\omega)N(A_0)/2}{H(\omega)L(\omega)N(A_0)/2+1} \tag{9}$$

In (10) A_0 denotes the limit cycle amplitude. Since the limit cycle oscillation fulfills the Barckhausen criterion and $H(\omega)$ has a higher gain for lower frequencies, the term $H(\omega)L(\omega)N(A_0)/2$ will always be larger than $1/2$ for frequencies lower than the limit cycle frequency. The input output relation thus will approximately be a unit gain relationship. The same calculation for the transfer of the distortion term $(d(V_{in}))$ to the output gives:

$$\frac{V_{out}}{d(V_{in})} = \frac{1}{H(\omega)L(\omega)N(A_0)/2+1} \tag{10}$$

This equation will tend to zero for high loop gain of the forward filter $H(\omega)$ and frequencies below the limit cycle frequency. Since an ADSL contains a large number of tones which are uncorrelated, we can approximate the contribution of distortion to the ADSL-signal as a white noise source σ_d. The output can thus be

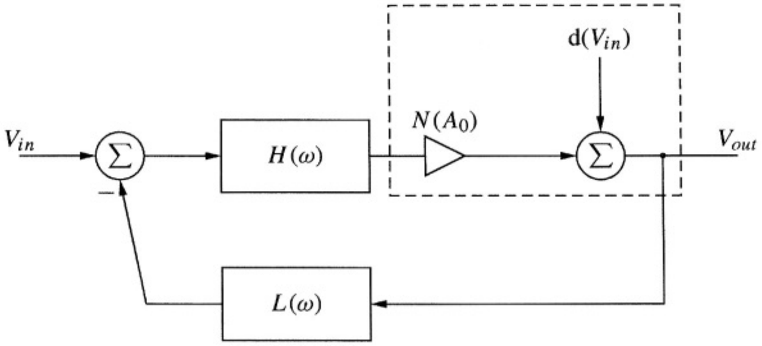


Figure 10: Simplified model of a SOPA, using a linearised model for the comparator (dashed box).

described as :

$$\begin{aligned} V_{out} &= \frac{H(\omega)N(A_0)/2}{H(\omega)L(\omega)N(A_0)/2+1}V_{in} + \frac{1}{H(\omega)L(\omega)N(A_0)/2+1}\sigma_d \\ &= \frac{H(\omega)}{H(\omega)L(\omega)+2\,H(\omega_0)L(\omega_0)}V_{in} \\ &\quad + \frac{2\,H(\omega_0)L(\omega_0)}{H(\omega)L(\omega)+2\,H(\omega_0)L(\omega_0)}\sigma_d \end{aligned} \tag{11}$$

Since the loop filter is low-pass, $L(\omega)$ can be approximated to 1. By stating $H(\omega)L(\omega) \gg 2\,H(\omega_0)L(\omega_0)$, (11) is approximated to :

$$V_{out} \approx V_{in} + \frac{H(\omega_0)}{H(\omega)}\,\sigma_d \tag{12}$$

The output distortion is thus shaped by the inverse of the forward filters. An integrator is thus an ideal forward filter for its high in-band gain, like noise-shaping is constructed in a $\Delta\Sigma$-converter. In the following we will define a n-th order SOPA as a SOPA with n integrators in the forward path.

5. Design Example

Figure 7 shows a realization of a zeroth order SOPA line driver in .25 µm CMOS technology. A third order loop filter is used to construct the limit cycle oscillation. The schematic is shown in figure 11(a). The different building blocks are grouped

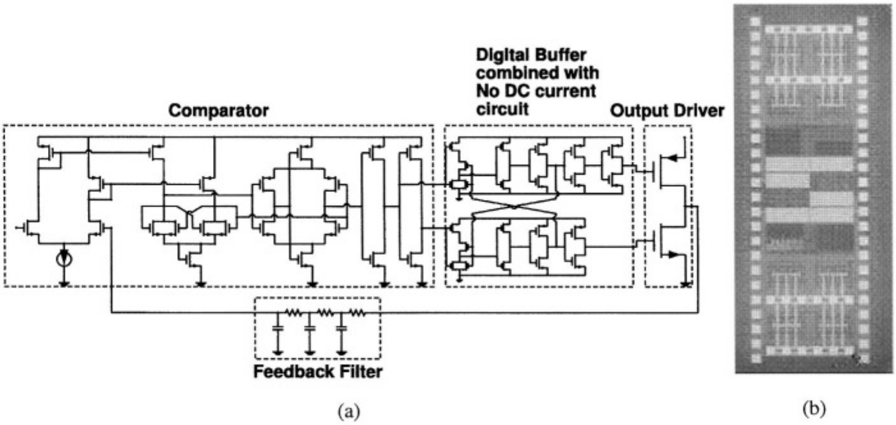


Figure 11: Circuit schematic (a) of the implemented SOPA line driver [21] and chip photograph (b).

corresponding to the block schematic of figure 7. It also reveals clearly the no shoot-through circuitry that is include to improve the output efficiency. Also the binary tree layout of the output inverter to obtain an equal current spread can be clearly observed from the chip photograph.

Figure 12 shows a Missing Tone Power Ratio (MTPR) measurement for the SOPA line driver of figure 11. For this test a DMT spectrum of 185 tones is applied to the line driver. Tones 50,100 and 150 are left out to collect the distortion power. The ratio between the output power and the power collected in these antenna tones is the MTPR and a good measure for the maximum bit-error rate that is achievable through the channel. From these measurements it becomes clear that the SOPA line driver is compliant with the ADSL G-Lite specifications and this for an efficiency of 41% which is much higher than what is theoretically achievable with a class AB or a two supply segmented class G power amplifier.

In order to reach ADSL and VDSL specifications, a third order SOPA power amplifier has been designed and processed [22, 23]. The output stage can be clearly seen. The drivers take more than one third of the overall chip area. The used integrators can also be clearly distinguished. The sizing towards the output is done be scaling the integrator capacitors. Since the non-linearity of an integrator is shaped by the preceding integrators. Therefor, the specifications of the first integrator will be the most stringent. In order to meet the linearity specifications the

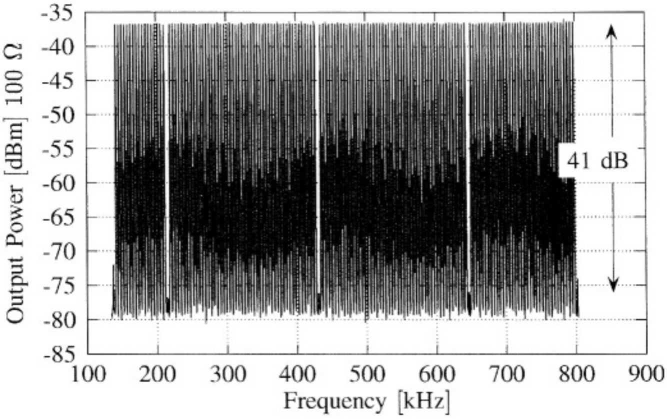


Figure 12: Measured output spectrum of the zeroth order SOPA when applying a DMT - signal with a CF of 5.6

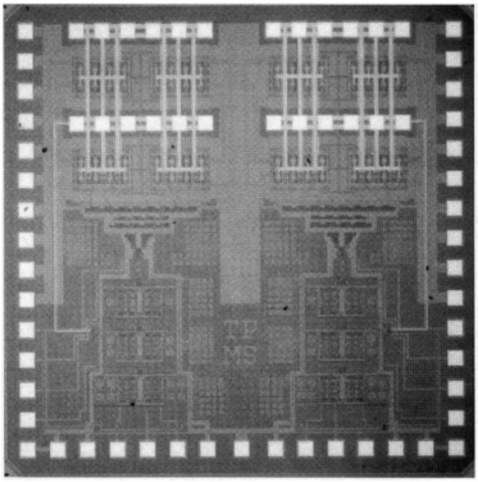


Figure 13: Chip photograph of the realised third order SOPA

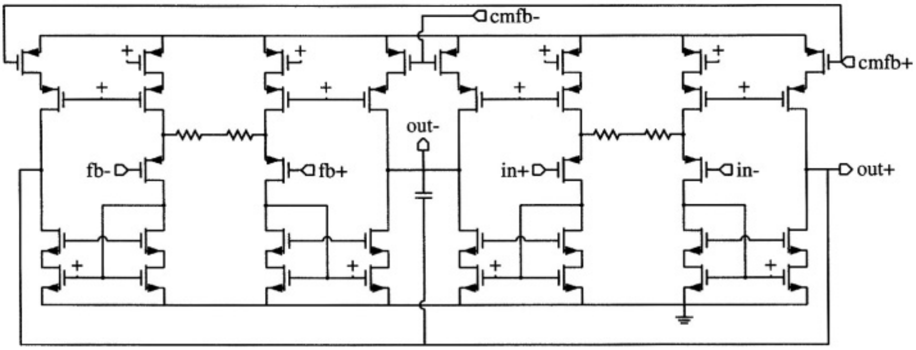


Figure 14: Schematic of the integrator used in the third order SOPA design.

resistive source degeneration has been used. Figure 14 shows the schematic used in the third order design. The integrator used, has four inputs to make a fully differential output. The increase in power supply rejection ratio is mandatory since the continuous time integrator operates in an environment of high power switching currents. The supply lines of the most critical analog building blocks are also decoupled as close as possible to the blocks themselves as can be observed from the chip photograph.

The linearity improvement by the resistive source degeneration is given by:

$$HD_3 = \frac{-1}{32} \frac{g_m^2}{I_{Bias}^2 (g_m R_E + 1)^3} \tag{13}$$

Since the integrator's unity gain frequency is given by:

$$f_0 = \frac{g_m}{2\pi (g_m R_E + 1) C_{int}}, \tag{14}$$

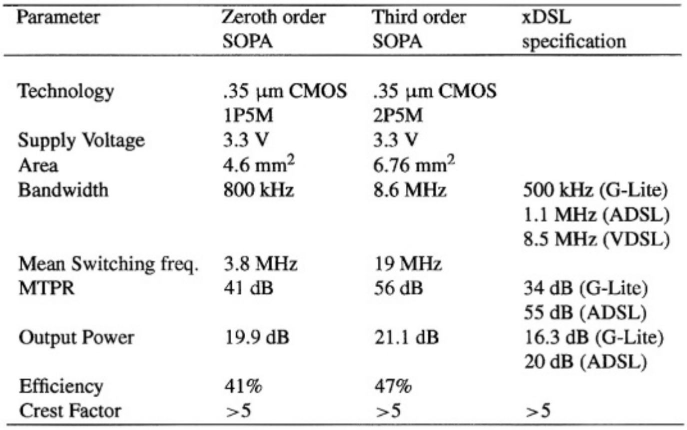
equations (13) and (14) together with the maximal allowed *kT/C*-noise will fully determine the integrator's sizing.

The measurement results of both line driver designs are summarized in table 2. From these the improvements by going to higher orders are clearly visible.

6. Conclusions

The class D power amplifier technique has been popular for decades in the design of highly efficient audio output amplifiers. Recently, due to the advent of novel

Table 2: Performance Summary of the two SOPA line drivers



Parameter	Zeroth order SOPA	Third order SOPA	xDSL specification
Technology	.35 μm CMOS 1P5M	.35 μm CMOS 2P5M	
Supply Voltage	3.3 V	3.3 V	
Area	4.6 mm^2	6.76 mm^2	
Bandwidth	800 kHz	8.6 MHz	500 kHz (G-Lite) 1.1 MHz (ADSL) 8.5 MHz (VDSL)
Mean Switching freq.	3.8 MHz	19 MHz	
MTPR	41 dB	56 dB	34 dB (G-Lite) 55 dB (ADSL)
Output Power	19.9 dB	21.1 dB	16.3 dB (G-Lite) 20 dB (ADSL)
Efficiency	41%	47%	
Crest Factor	>5	>5	>5

modulation techniques like DMT, these type of switching power amplifiers reappear in the general research picture. Due to their high efficiency, which is barely dependent on the high CF of those modulations, class D line drivers are the key component for high density central office xDSL boards.

An important family of class D line drivers, the self oscillating power amplifiers has been presented. The important phenomenon of oscillator pulling towards synchronization improves the bandwidth by a factor of 2 while the output filter specifications are heavily relaxed.

The two presented SOPA line drivers prove that even for high demanding systems like xDSL, it is possible to design compliant self oscillating class D line drivers in a submicron CMOS technology.

References

[1] R. van der Zee, “High efficiency audio power amplifiers, design and practical use,” Ph.D. dissertation, Universiteit Twente, 1999.

[2] D. Macq, “DMT ADSL circuits and systems,” in *ISSCC : xDSL Broadband Interactive Communications,* Feb. 1998.

[3] B. Schweber, “Analog front ends bridge the xDSL-to-real-world chasm,” *EDN magazine,* pp. 48–64, Apr. 1999.

[4] K. Maxwell, “Asymmetric digitil subscriber line : Interim technology for the next forty years,” *IEEE Commun. Mag.,* pp. 100–106, Oct. 1996.

[5] L. Cloetens, “Broadband access : The last mile,” in *1SSCC Digest of Technical Papers,* Feb. 2001, pp. 18–21.

[6] R. Apfel, “ADSL transceivers : Optimizing power, cost, function and performance,” in *ISSCC: SSCTC Workshop on Analog Telecom Access Circuits and Concepts,* Feb. 2002.

[7] R. Benton, *et al.,* “A high-voltage line driver (HVLDR) for combined voice and data services,” in *ISSCC Digest of Technical Papers,* Feb. 2001, pp. 302–303.

[8] D. Rossi, “Silicon technologies and circuit topologies for high efficiency in ADSL line drivers,” in *ISSCC: SSCTC Workshop on Analog Telecom Access Circuits and Concepts,* Feb. 2002.

[9] M. Cresi, *et al.,* “An ADSL central office analog front-end integrating actively-terminated line driver, receiver and filters,” in *ISSCC Digest of Technical Papers,* Feb. 2001, pp. 304–305.

[10] M. Ingels, S. Bojja, and P. Wouters, “A $0.5\ \mu$m CMOS low-distortion low-power line driver with embeded digital adaptive bias algorithm for integrated ADSL analog front-ends,” in *ISSCC Digest of Technical Papers,* Feb. 2002, pp. 324–325.

[11] F. Sabouti and R. Shariatdoust, “A 740 mW ADSL line driver for central office with 75 dB MTPR,” in *ISSCC Digest of Technical Papers,* Feb. 2002, pp. 322–323.

[12] J. Sevenhans, *et al.,* “Driving the DSL highway : high speed, high density, low power, low cost,” in *Proceedings European Solid-State Circuits Conference,* Sept. 2002, pp. 555–562.

[13] J. Pierdomenico, S. Wurcer, and B. Day, “A 744 mW adaptive supply full-rate ADSL CO driver,” in *ISSCC Digest of Technical Papers,* Feb. 2002, pp. 320–321.

[14] I. Kalet, “The multitone channel,” *IEEE Trans. Commun.,* vol. 37, no. 2, pp. 119–124, Feb. 1989.

[15] D. MestDagh, P. Spruyt, and B. Biran, “Effect of amplitude clipping in DMT-ADSL transceivers,” *Electronics letters,* vol. 29, no. 15, pp. 1354–1355, July 1993.

[16] T. Sowlati and D. Leenaerts, “A 2.4 GHz $0.18\mu m$ CMOS self-biased cascode power amplifier with 23 dBm output power,” in *ISSCC Digest of Technical Papers.* IEEE Press, Feb. 2002, pp. 294–295.

[17] E. Dallago, “Advances in high-frequency power conversion by delta-sigma modulation,” *IEEE Trans. Circuits Syst. I,* vol. 44, pp. 712–721, Aug. 1997.

[18] K. Philips, J. van den Homberg, and E. Dijkmans, “PowerDAC : a single chip audio DAC with a 70% efficient powerstage in $0.5\mu m$ - cmos,” in *ISSCC Digest of Technical Papers.* IEEE, 1999.

[19] A. Gelb and W. Vander Velde, *Multiple-input Describing Functions and Nonlinear System Design.* McGraw-Hill Book Company, 1968.

[20] E. Roza, “Analog-to-digital conversion via duty-cycle modulation,” *IEEE Trans. Circuits Syst. II,* vol. 44, no. 11, pp. 907–914, Nov. 1997.

[21] T. Piessens and M. Steyaert, “SOPA : A high-efficiency line driver in $0.35\mu m$ CMOS using a self-oscillating power amplifier,” in *ISSCC Digest of Technical Papers,* Feb. 2001, pp. 306–307.

[22] ——, “A central officed combined ADSL-VDSL line driver in $.35\ \mu m$ CMOS,” in *Proceedings Custom Integrated Circuits Conference.* IEEE, May 2002, pp. 45–48.

[23] ——, “Highly efficient xDSL line drivers in $0.35\ \mu m$ CMOS using a self-oscillating power amplifier,” *IEEE J. Solid-State Circuits,* vol. 38, no. 1, pp. 22–29, Jan. 2003.

This page intentionally left blank

Class G/H Line Drivers for xDSL

John Pierdomenico
Analog Devices, Inc. Wilmington, MA

Abstract

The limitations of Class AB topologies for xDSL applications are presented. A general discussion of Class G/H/H-like amplifier topologies for xDSL applications follows during which a theoretical power consumption derivation of these systems driving discrete multi-tone (DMT) waveforms is presented. From this discussion, circuit examples of each of the three topologies are presented processing identical signals and delivering identical powers for total power consumption comparisons. Finally, several measurements of the AD8393 (Class H-like) *ALP amplifier*^{TM} system are given.

1. Introduction

The discrete multi-tone (DMT) modulation employed in ADSL systems creates a waveform with high peak to rms levels (crest factor or CF) that poses severe challenges for the design of power efficient, high output current line drivers. Central Office applications require that an average of 20.4dBm(110mW or $3.31V_{rms}$ differential) be delivered to a 100Ω phone line, but peak signals can require a 36x (3.9W or 19.86Vp differential) increase in instantaneous power. Supply voltages on many Class AB drivers are typically set to +/-12V to handle peak signals resulting in total supply power consumption, TSPC, in the range of 1W. This TSPC includes IC power dissipation, line power dissipation, transformer and back termination power losses.

Reducing the power consumption, regulated power supplies, and overall component count of these line driving systems has become a focal point as line card densities go beyond 48 ports. One recently reported single 15V DC supply driver delivered 20.4dBm of line power for 1.1 W of TSPC [1], while another dual supply, +/-12V, driver achieved this goal for 740mW of TSPC and had all of the resistive gain network integrated on chip [2]. In an effort to remove all regulated supplies, one fully integrated interface (analog front end and line driver), which included the line driver function, achieved full rate line power for

A. van Roermund et al. (eds.), Analog Circuit Design, 333-358.
© 2003 *Kluwer Academic Publishers. Printed in the Netherlands.*

a total power consumption of 2.6W (of which the line driver DC current component was 10mA on an unregulated –48V supply) [3].

Several power reduction techniques exist to optimize the efficiency of ADSL driver systems. Many designs now incorporate what is referred to as active line termination [4]. An example of this is illustrated with figures 1 and 2.

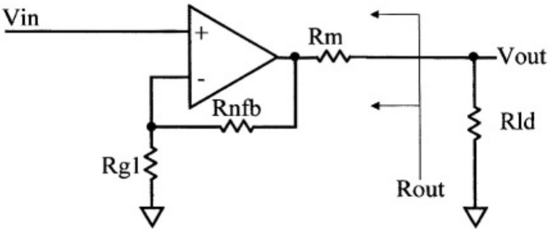


Figure 1. Active Termination

Figure 1 is an amplifier application whose gain and output impedance can be expressed as:

$$Av = \frac{Vout}{Vin} = \left(\frac{Rg1 + Rnfb}{Rg1}\right) * \left(\frac{Rld}{Rld + Rm}\right) \qquad (1)$$

$$Rout = Rm \qquad (2)$$

In an xDSL application, this circuit will be driving a transformer and it's output impedance will be required to match that of a reflected load thru the transformer.

The price paid for this is a 2 to 1 drop in voltage from the output of the amplifier to the load. To accommodate this voltage drop, the amplifier supply voltage must be increased. This increases the power consumption of the circuit.

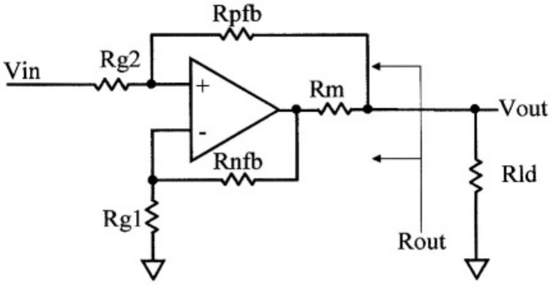


Figure 2. Active Termination

Figure 2 shows the same amplifier with an active termination to generate the output impedance. The gain and output impedance of this circuit can be expressed as:

$$Av = \frac{Vout}{Vin} = \frac{Rld * Rpfb * (Rnfb + Rg1)}{Rg1 * (Rg2 + Rnfb) * (Rld + Rm) - Rg2 * Rld * (Rnfb + Rg1)} \tag{3}$$

$$Rout = \frac{Rm}{1 - (1 + \frac{Rnfb}{Rg1}) * (\frac{Rg2}{Rpfb + Rg2})} \tag{4}$$

This technique provides the same output voltage to a load on lower supply voltages while still providing proper return signal termination through a combination of positive and negative feedback. The drop across resistor Rm can be as low as 0.1x that of the load resistance drop. Efficiency is improved but a penalty is paid in the form of return signal attenuation and resultant signal to noise ratio degradation.

While active line termination reduces the power loss due to back termination, thus improving the system efficiency, it does nothing to improve the driver efficiency. For a standard Class AB amplifier using active line termination, the 1W TSPC can improve to around 750mW.

Another technique gaining attention is the use of multi-supply (Class G) topologies utilizing 2 sets of power supplies. This system turns on a higher supply voltage when the output signal surpasses a threshold level. The intent is to operate the driver from lower supply voltages most of the time resulting in reduced TSPC. Active line termination has been used in these systems and TSPC values have been reported down to 695mW on +/-5V and +/-15V supplies, and 610mW on +/-4V and +/-8V supplies [5,6].

Similar to the Class G system is the Class H/H-like amplifier system. Typically these systems have supplies that vary linearly based on analog signal levels. With this technique, the improved efficiency of Class G can be obtained without the need for a secondary supply. Class H-like amplifiers having TSPC's in the range of 600-700mW have recently been reported [7,8].

Note that throughout this discussion, references to Class H/H-like will be made. This is done as there are several amplifiers claimed as Class H [9, 10] that are slightly different than the system topology to be discussed. The author would like to note that several references to Class H systems have slightly different pretenses and thru discussions with others, has come to the conclusion that the Class H terminology was designated to differentiate it from Class G and that there seems to be no unified definition for Class H.

2. Efficiency

Before covering circuit operation of Class G/H/H-like topologies, it is instructive to analyze the theoretical efficiencies of traditional Class AB, Class G and Class H systems when processing a DMT signal. This analysis helps generate the motivation for applying Class G and Class H amplifiers to xDSL systems. Figure 3 represents a typical DMT waveform in the time domain. The crest factor (CF) of the waveform is 5.3, and assuming a 100Ω load, the power is 20.4dBm. What is unique about this waveform is the aperiodicity of its peak. This waveform results from the combination of 224 QAM like signals from 130kHz to 1.1Mhz spaced at intervals of 4.3125kHz. The addition of these QAM signals creates a signal with an approximate Gaussian amplitude distribution. The peak occurs when many of the channels align in phase.

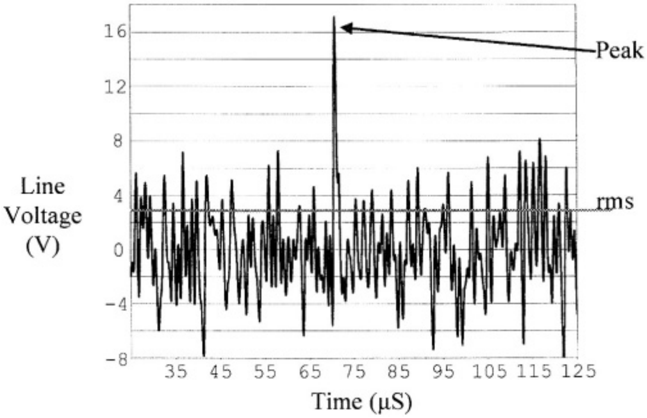
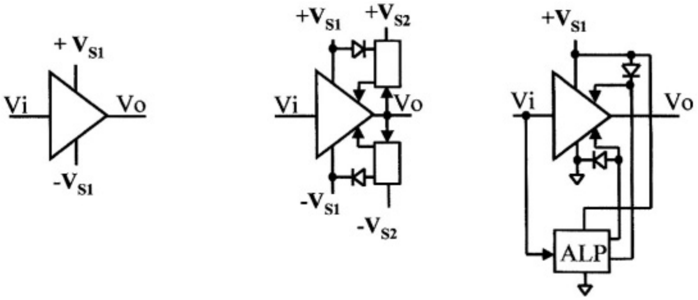


Figure 3. Typical ADSL DMT Line Waveform

Figure 4 shows three different ideal line driver systems: Class AB, Class G and Class H-like. While the diagram shows single ended operation, for DSL, all drivers operate in a differential configuration. Additionally for this exercise, it is assumed that there are no losses due to coupling or back termination.



Class AB

Class G

Class*
H/H-like

*ADI Patent Pending

Figure 4. Ideal Amplifier Supply Comparison

With reference to figure 3, Class AB system supply voltages are set to handle peak signal voltages at all times. For an rms level of 3.3V and a CF of 5.3, the supplies of an ideal Class AB system can be set to no lower than V_{S1}=+/-8.75V ([5.3*3.3]/2). In the ideal Class G system, the primary supply voltages are set to process the rms level of 3.3V (V_{S1}=+/-1.65V) while the secondary supply voltages are set to the peak level of V_{S2}=+/-8.75V. Switching between supply voltages can be based on input or output signal levels, or from an external signal. In the ideal Class H/H-like system, the supply voltage is set to process the rms level with a single V_{S1}=5.83V supply. Input sampling circuitry boosts internal supply voltages to the required level (5.83V*3=17.49V) to process peak signals at the amplifier's output.

The theoretical efficiency of an ideal Class G and Class H/H-like system has been derived [7] and can be restated (assuming zero amplifier quiescent power, P_Q) in a practical manner as:

$$\varepsilon(ClassG) = \frac{1}{A\sqrt{\frac{2}{\pi}}(1-e^{\frac{-A^2}{2}}) + B\sqrt{\frac{2}{\pi}}e^{\frac{-A^2}{2}}} \times 100\,\% \qquad (5)$$

$$\varepsilon(ClassH) = \frac{1}{A\sqrt{\frac{2}{\pi}}(1-e^{\frac{-A^2}{2}}) + 2\cdot A\sqrt{\frac{2}{\pi}}e^{\frac{-A^2}{2}}} \times 100\,\% \qquad (6)$$

where A is the number of standard deviations above the rms signal level for the primary supply and is defined as:

$$A = \frac{V_{S1}}{V_{Orms}} \qquad (7)$$

And B is similar but related to the secondary supply, V_{S2}, present in Class G systems and is defined as:

$$B = \frac{V_{S2}}{V_{Orms}} \geq CF \qquad (8)$$

Note that for headroom constraints in the Class H/H-like system:

$$A \geq \frac{CF}{3} \qquad (9)$$

For example, to obtain the maximum efficiency in an ideal differential Class G system where V_{orms}=3.31V, A=2, and the CF=5.3; the primary supply, V_{S1}, would be set to +7-3.3V while the secondary supply, V_{S2}, is +/-8.75V. The ideal maximum efficiency for this condition would be 51%.

For an ideal Class H/H-like amplifier system, the ideal maximum efficiency for this same condition would be 58%.

Note that, for Class G and H/H-like systems, as the primary supply voltage, V_{S1}, increases, the following begins to occur:

The second denominator term in equations (5) and (6) becomes less significant as power consumption in the primary supply begins to dominate. Additionally, as the primary supply voltage approaches the minimum required to process the CF in a Class AB system (+/-8.75V), the efficiencies (5 & 6) become that of a Class AB amplifier, which can be shown to be:

$$\varepsilon = \frac{1}{\frac{V_S}{V_{Orms}}\sqrt{\frac{2}{\pi}}} \times 100\ \% \qquad (10)$$

Under the same CF conditions above, the ideal efficiency for a Class AB amplifier would be 23.6% at a supply voltage of +/-8.75V.

Figure 5 plots the ideal efficiency vs. primary supply voltage, V_{S1}, for the above efficiency equations (5), (6) & (10). It compares the three ideal amplifier topologies processing a DMT signal. Again, the CF is 5.3, the rms line voltage is 3.3V (20.4dBm into 100Ω) and it has been assumed that there are no losses due to back termination, line coupling, diode drops, amplifier headroom, etc., for all systems. Also, the switching point for the Class G and Class H systems was set equivalent to A, eliminating headroom loss and maximizing efficiency for both systems.

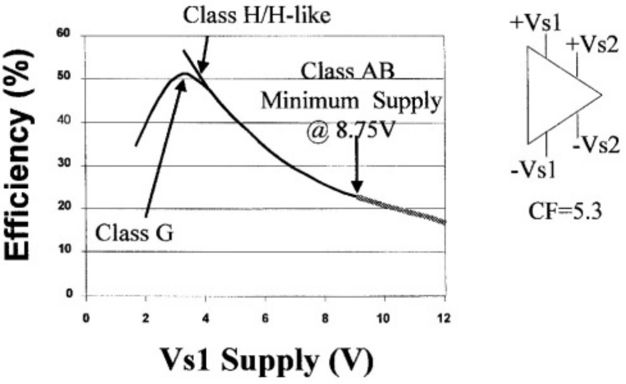


Figure 5. Ideal Efficiency vs. Primary Supply
Vs2=8.75V; [CF*(Vorms)/2]

The graph is illustrative in a number of important ways:

First, an ideal Class AB system can only operate at minimum supplies of +/-8.75V; $(CF^*V_{Orms})/2$. The best efficiency an ideal Class AB system achieves for this supply is 23.6%. This equates to an ideal TSPC of 464mW. In practice, this is not realizable as supply headroom and the amplifier quiescent power P_Q, will increase power dissipation. Additionally, losses due to back termination and coupling add power to the TSPC. Typically, Class AB topologies today are operating on supply voltages of +/-12V. The ideal efficiency at this supply level is 17.2%. Again, assuming no amplifier P_Q, TSPC is 633mW. If 5mA of amplifier quiescent current is added, TSPC rises to approximately 753mW, the minimum total supply power consumed a Class AB amplifier can achieve when driving 20.4dBm into 100Ω with a CF of 5.3; assuming no back termination or coupling losses.

Second, for the Class G and Class H/H-like system, the ideal efficiency reaches 40% as the total supply voltage is dropped to 10V (Vs1=5V). For the Class G topology, the ideal efficiency will approach a maximum of 51% as total supply drops to 6.6V. For the Class H/H-like topology, the ideal efficiency approaches 55% for this same total supply voltage. For both systems, this is of little practical use, as headroom requirements will usually dictate the use of higher supply voltages.

Additionally, both systems will have losses due to back terminations, coupling and diode drops. However, these systems can be operated on a total supply range of 10-12V setting the ideal efficiency range at 35-40%. If the amplifier Iq is set to 7mA in the above example, on 12V of total supply, the Class G/H/H-like amplifier systems can generate the same 20.4dBm to the line on less than 425mW of TSPC (assuming no other losses).

As mentioned above, there will be practical limitations to obtaining the theoretical best efficiency with all systems.

3. Class G topology

Figure 6 is a simplified schematic of a generic single ended Class G system. The schematic shows primarily the operation of the output stage of a Class G system. When the circuit is processing rms voltage levels, current drawn from the supply labelled Vorms/2, passes thru diode D1 and transistor Q1 and into the load. During this operation, transistor Q2 is off and no current is drawn from the supply labelled CF*Vorms/2. When a peak occurs, the voltage at node Vout increases to the point where transistor Q2 now becomes active (as the base of transistor Q2 increases with Vout via Z1). During this operation, diode D1 becomes reverse biased and current is drawn from the supply labelled CF*Vorms/2. The current path now consists of transistors Q1 & Q2 and is passed into the load. As peak signals occur rarely [11], minimal current is drawn from the higher supply voltage and overall IC power dissipation is minimized.

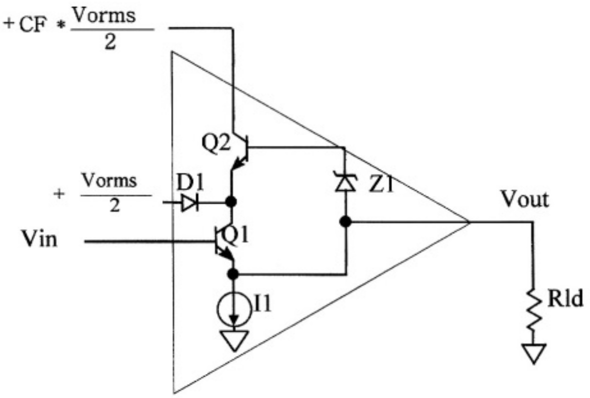


Figure 6. Class G Operation

When designing this type of amplifier there are several design challenges that should be addressed:

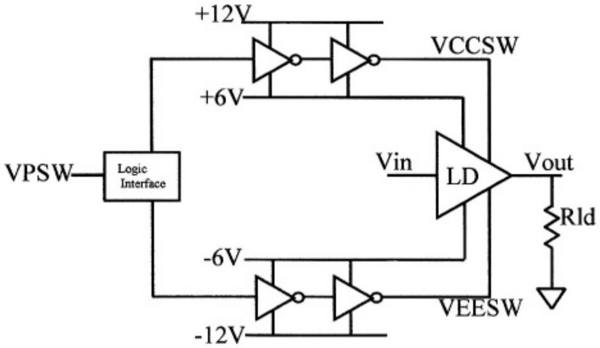
This topology requires an extra set of supplies. The extra hardware/supply busing that is required may limit the circuit's ability in increase the number of ports per board.

The voltage drop thru zener Z1 must be accounted for in the high current supply, which adds power consumption. This can be overcome by a zero overhead approach to be discussed later.

There are timing complexities that must be dealt with. The supplies can be triggered either internally via the incoming signal or externally with a control signal.

Transistor Q2 ultimately must handle the same peak current as Q1, whose area is sized to handle the large peak currents. This increase in area will add cost to the final solution.

As mentioned above, the zero overhead approach is one way to overcome the large voltage drop in the above design [6]. The functional operation of this approach is shown in figure 7. Again, there are two sets of supplies, but it is how these two supplies are presented to the amplifier that is unique. The signal that controls the supply that powers the output stage is VPSW. When the amplifier output signal level is predicted to overcome the lower supply voltages, VPSW will switch to the higher supply as shown in figure 8. When done so, the higher supply is passed thru the buffers to power up the output stage during the peak signal event. This architecture operates without the need for large voltage drop zener diodes.



Source: ISSCC2003
610mW Zero Overhead Class G Full Rate ADSL CO Line Driver

Figure 7. Class G Zero Overhead

Source: ISSCC2003
610mW Zero Overhead Class G Full Rate ADSL CO Line Driver

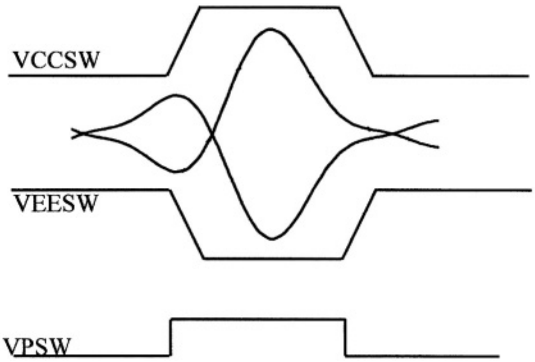


Figure 8. Class G Zero Overhead

Again, there are several design challenges/issues that should be addressed:

This topology still requires an extra set of supplies. The extra hardware/supply busing that is required may limit the circuit's ability in increase the number of ports per board.

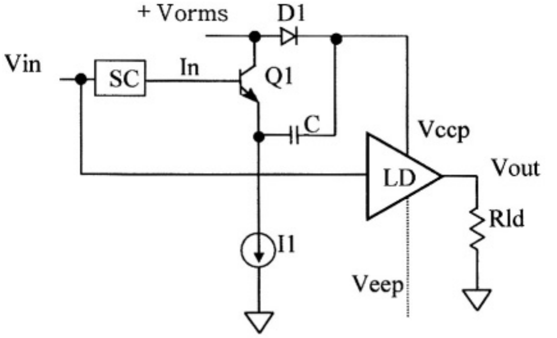
Peak prediction of the incoming signal. It is important to note that the peak signal is the phase sum of the available channel signals. This level can be altered by board/system parasitics that the predictor may not be aware of, thus causing potential mis-switching of the supplies. This architecture will work best if AFE and line driver are on the same piece of silicon where parasitics can be controlled.

The switches that the high supply passes thru must have low resistance as they will also be passing peak current. This adds to the die area.

With these switches, shoot thru currents and edge coupling must be dealt with and eliminated.

4. Class H topology

Figure 9 is a simplified schematic of a generic single-ended Class H/H-like system. When a large peak voltage signal reaches node IN, the emitter follower, Q1, along with diode, D1, and capacitor C1, drives the voltage at node VCCP above VCC in the shape of an amplified version of the input signal. Peak current flows through transistor Q1 and capacitor C1 on its way to the load. Diode D1 is reverse biased providing isolation to the DC supply applied to VCC. The threshold voltage for activation is set by a combination of gain in the signal conditioning path and the time-averaged voltage across the capacitor C1. Some of this occurs in the block labeled SC. Below this threshold voltage, the load current flows through diode D1 and into the driver amplifier. Current source, I1, recharges the capacitor in time for the next peak.



*ADI Patent Pending

Figure 9. Class H/H-like Operation

For ADSL DMT signals, given a probability of clipping, the interval between peaks can be determined for a frame length and total number of samples per frame. Assuming ideal Gaussian statistics, the probability that a peak occurs varies from once every 4.6s for a 5.3CF up to once every 7.6min. for a 6.4CF [11]. Since the peak appears infrequently, the current I1 and capacitor C1 are relatively small as the recharge rate is much shorter than the time between voltage peaks. Since the DC supplies are set to handle the rms level plus headroom, power reduction is substantial as the internal voltage supplies only increase on peak signals.

While this topology is more eloquent when compare to the two Class G approaches above, it is not without design challenges/issues. These include:

Extra diodes and capacitors for proper operation. While requiring less board area than an extra supply, these still are of concern. However, a process with a good Schottky diode can be used to limit the extra components. This will add slightly to die area.

There are timing complexities that must be addressed when firing the pumps. The pumps can either be activated via an external signal or from the incoming signal waveform.

Transistor Q1 will handle the same peak current as the output transistor, therefore it must be sized appropriately. This will add to die area.

While each of the above three examples will save power over the Class AB system, they do so for a slight cost adder. Eventually, end users will determine the value of this power savings and decide their worthiness for ADSL applications.

5. Circuit examples

It is instructive to compare the three amplifier classifications (AB, G and H) while driving similar signals under similar DC operating conditions. Figures 10 and 11 shows a Class AB amplifier in a non-inverting gain of about 19 without active termination. Figures 12 and 13 do the same for Class G while figures 14 and 15 presents a Class H/H-like amplifier. Note that the Iamp circuit block in figure 14 is the same as figure 11.

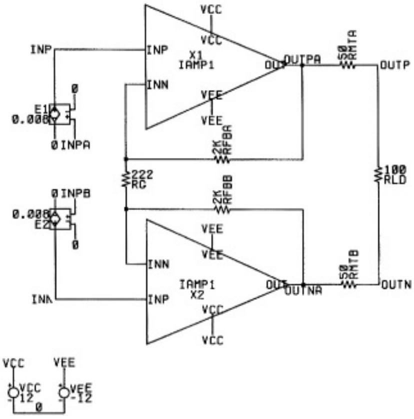

VCC
VCC
INP
INP
X1
IAMP1
OUT
OUTPA
50
RMTA
OUTP
INN
0
E1
0.008
VEE
VEE
0 INPA
2K
RFBA
222
RC
100
RLD
0 INPB
2K
RFBB
0.008
E2
0
VEE
VEE
INN
IAMP1
X2
OUT
OUTNA
50
RMTB
OUTN
INN
INP
VCC
VCC
VCC
VEE
VCC
12
VEE
-12
0

Figure 10. Class AB Amplifier

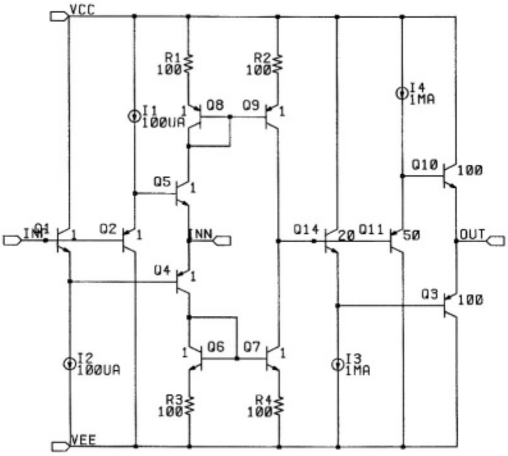

VCC
R1
100
R2
100
I4
1MA
I1
100UA
1
Q8
Q9
1
Q10
100
Q5
1
INP
Q1
1
Q2
1
INN
Q14
20
Q11
50
OUT
Q4
1
Q3
100
1
Q6
Q7
1
I2
100UA
I3
1MA
R3
100
R4
100
VEE

Figure11. Class AB Amplifier

The Class AB amplifier in figure 11 is a standard current feedback block. The input buffer (Q1, Q2, Q4 & Q5) mirrors the signal to the high impedance node (base of Q11, Q14). From there, the signal is buffered to the output via transistors

Q3, Q10, Q11 & Q14. This block is designed to drive over 300mA of peak current.

The Class G amplifier in figure 13 is very similar to the Class AB amplifier in figure 11. It has been fitted with a couple of extra devices (Q12 & Q13) and diodes (D1 & D2) to make Class G operation possible. During small signals, the amplifier is powered thru diodes D1 & D2. During peak signals, transistors Q12 & Q13 will activate pulling the internal supply rails above VCCL/VEEL. During this time, the diodes, D1 & D2, are reverse biased and the amplifier is powered by VCCH/VEEH thru devices Q12 & Q13. As such, devices Q12 & Q13 must be sized similarly to devices Q3 & Q10 if proper distortion is to be maintained.

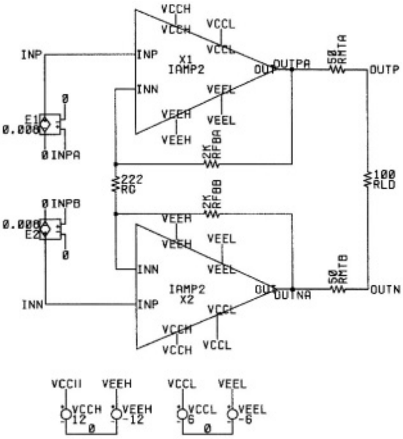

VCCH
VCCH
VCCL
VCCL
INP
INP
X1
IAMP2
OUT
OUTPA
50
RMTA
OUTP
INN
VEEL
VEEH
VEEL
VEEH
0
E1
0.000
0 INPA
2K
RFBA
222
RG
100
RLD
2K
RFBB
0 INPB
0.000
E2
0
VEEH
VEEH
VEEL
VEEL
INN
IAMP2
X2
OUT
OUTNA
50
RMTB
OUTN
INN
INP
VCCL
VCCH
VCCH
VCCL
VCCH
VEEH
VCCL
VEEL
VCCH
12
VEEH
-12
VCCL
6
VEEL
-6
0
0

Figure 12. Class G Amplifier

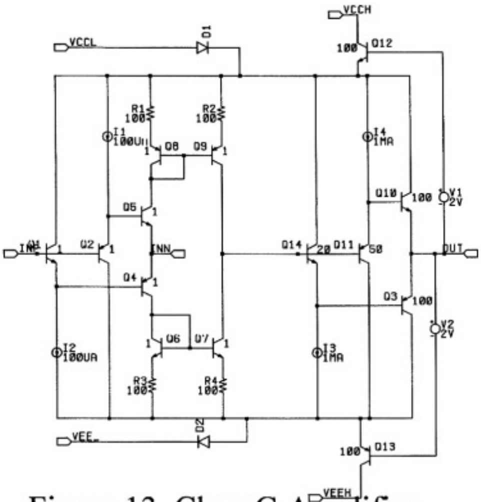

VCCH
100
Q12
VCCL
D1
R1
100
R2
100
I1
100U
I4
1MA
1
Q8
Q9
1
Q10
100
V1
2V
Q5
1
IN
Q1
1
Q2
1
INN
Q14
20
D11
50
OUT
Q4
1
Q3
100
V2
2V
1
Q6
Q7
1
I2
100UA
I3
1MA
R3
100
R4
100
D2
VEE
100
Q13
VEEH

Figure 13. Class G Amplifier

The Class H-like amplifier in figure 14 contains both the Class AB amplifier of figure 11 along with an *Adaptive Linear Power*^{TM} block shown in figure 15. During small signals, the amplifier is supplied by VCC and VEE thru diodes D1 & D2. When the incoming signal becomes large enough to trigger the *ALP amplifier*^{TM}, signals at nodes VCCP & VEEP increase to reverse bias diodes D1 & D2. During this time, peak current passes thru transistors Q1 & Q2 in figure 15. Again, since these devices are passing peak currents, they need to be sized similarly to the output devices of the main amplifier if proper distortion is to be maintained.

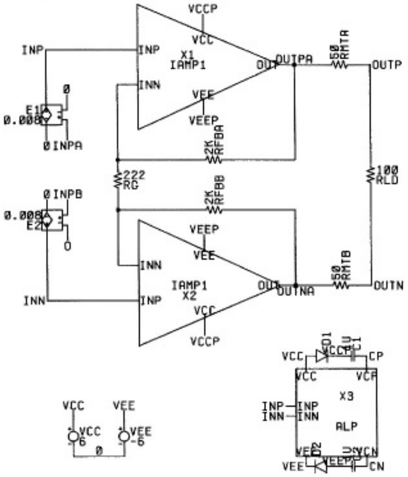

VCCP
VCC
INP
INP
X1
IAMP1
OUT
OUTPA
50
RMTA
OUTP
INN
0
E1
0.008
VEE
VEEP
0 INPA
2K
RFBA
222
RG
100
RLD
0 INPB
2K
RFBB
0.008
E2
0
VEEP
VEE
INN
IAMP1
X2
OUT
OUTNA
50
RMTB
OUTN
INN
INP
VCC
VCCP
D1
VCCP
1U
C1
VCC
CP
VCC
VCP
X3
INP
INN
INP
INN
ALP
VEE
D2
VEEP
1U
C2
VCN
VEE
CN
VCC
VEE
VCC
6
VEE
-6
0

Figure 14. Class H/H-like Amplifier

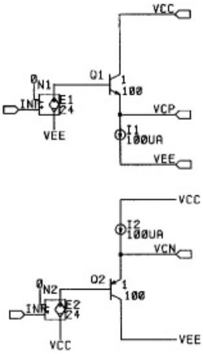

VCC
Q1
1
100
0
N1
INP
E1
24
VCP
VEE
I1
100UA
VEE
VCC
I2
100UA
VCN
Q2
1
100
0
N2
INN
E2
24
VCC
VEE

Figure 15. Class H/H-like Amplifier

All three topologies will be processing the same incoming level of DMT signal and with the same gain, will have identical signal levels at their outputs. The Class AB amplifier will be biased with +/-12V of DC supply. The Class G amplifier will have both +/-6V and +/-12V DC supplies. The Class H/H-like will have +/-6V DC supplies. All three will have a DC quiescent current of about 10mA. This DC current level is fairly reasonable considering that all three will be processing the same level of output signal. Additionally, the main amplifier topology is essentially the same for all three examples. What is slightly optimistic for the Class G/H/H-like is that the extra current needed to run peripheral circuitry is not included. However, these examples serve as a good starting point to help one better understand the operation of these different types of amplifiers and appreciate their power savings. Each amplifier will put out about 13dBm of line power and have similar MTPR distortion properties.

Figure 16 is the time domain plot of the Class AB amplifier. The supplies are more than adequate to handle the signal swing. However, as this peak occurs infrequently, this extra headroom is not necessary and power is wasted. Figure 17 is the time domain plot of the Class G amplifier. The lower supplies are shown in blue and are clearly not large enough to handle the peak output signal swing. The higher supplies are shown in red and current is drawn thru these when the output signal gets large enough. Figure 18 is the time domain plot of the Class H amplifier. Note that while the amplifier output surpasses the lower supplies, the pumping supplies support a much larger voltage as evident from the plot.

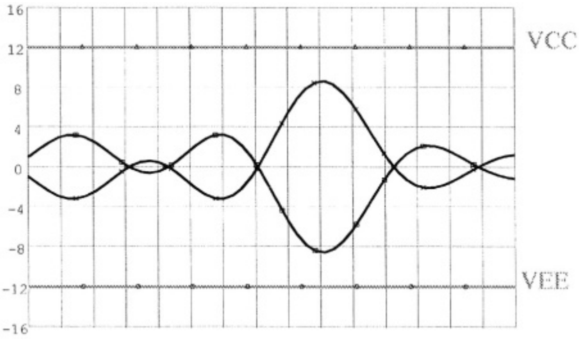


Figure 16. Class AB Time Domain

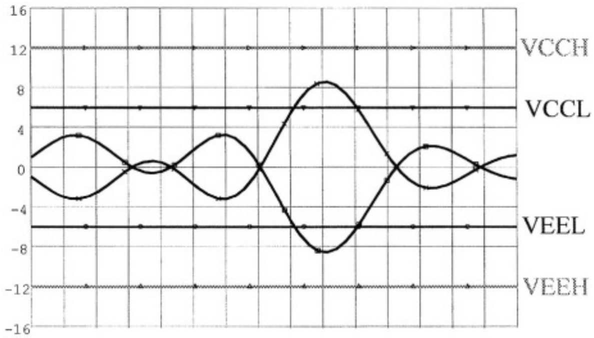


Figure 17. Class G Time Domain

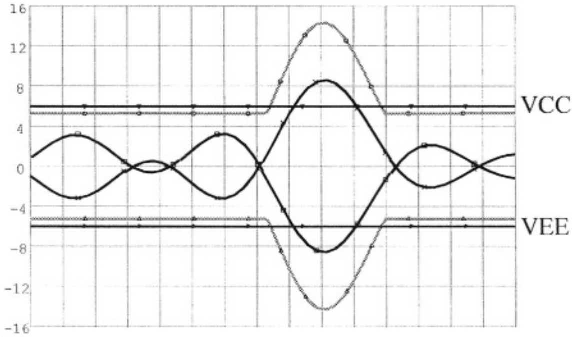


Figure 18. Class H Time Domain

The rms output levels of each circuit are show in figure 19. All three amplifiers are processing the same signal. Figure 20 shows the power dissipations of each amplifier system. Note that while the Class AB amplifier TSPC is slightly more than 450mW, the Class G and Class H/H-like is slightly less than 250mW.

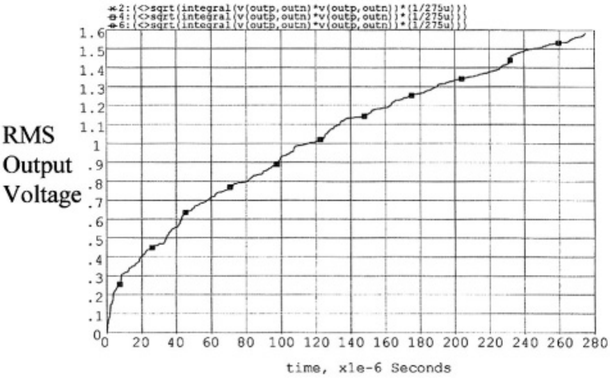


Figure 19. Output RMS Levels

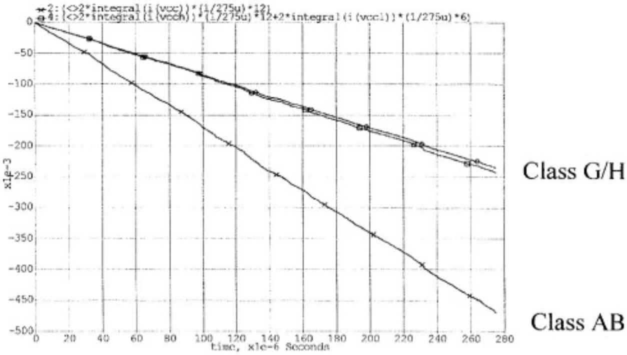


Figure 20. Total Power Consumption

6. Class H-like (AD8393) system measurements

One type of Class H-like system is Analog Devices' AD8393. This system will generate 24V of AC power supply swing on only 12V DC of supply.

Figure 21 is a scope photo of the driver amplifier output voltages vs. time along with the internal pumped voltages (VCCP & VEEP) of the *AD8393* and external fixed supplies (VCC & GND). The photo shows VCCP and VEEP tracking the output signal and exceeding the magnitude of the fixed external supplies by 6V above VCC and below GND. This 24V AC swing on the VCCP/VEEP bus permits the outputs of the amplifiers, VOUTP & VOUTN, to swing 18V on the 12V DC supply when the high input crest appears.

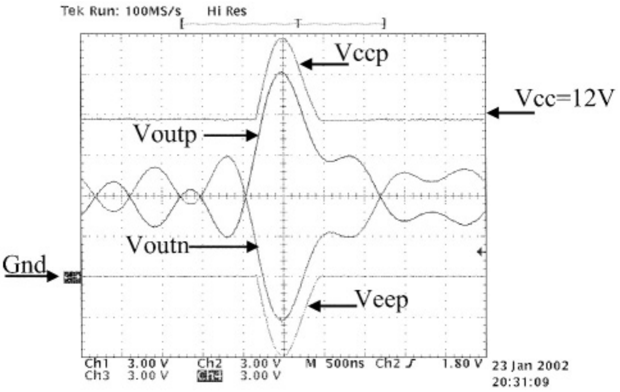


Figure 21. AD8393 Time Domain Measurement

Figure 22 is the MTPR measurement for this waveform in a 83Ω resistive load application (100Ω reflected thru the 1:1.1 transformer). This is equivalent to driving 20.4dBm into 100Ω. At 800kHz, MTPR was better than –70dBc. The TSPC from this plot was 650mW.

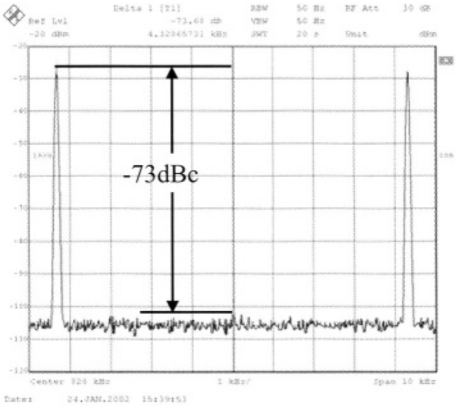
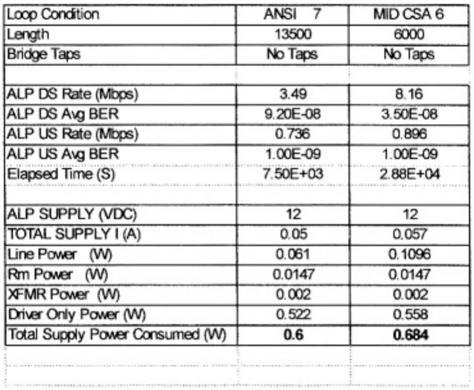


Figure 22. AD8393 MTPR Measurement

Figure 23 are the results of this amplifier cut into an existing reference design that was optimized for a Class AB amplifier. For the two line conditions presented, the TSPC ranges from 600-684mW while maintaining better than 1E-7 BER both upstream and downstream.



Loop Condition	ANSI 7	MID CSA 6
Length	13500	6000
Bridge Taps	No Taps	No Taps
ALP DS Rate (Mbps)	3.49	8.16
ALP DS Avg BER	9.20E-08	3.50E-08
ALP US Rate (Mbps)	0.736	0.896
ALP US Avg BER	1.00E-09	1.00E-09
Elapsed Time (S)	7.50E+03	2.88E+04
ALP SUPPLY (VDC)	12	12
TOTAL SUPPLY I (A)	0.05	0.057
Line Power (W)	0.061	0.1096
Rm Power (W)	0.0147	0.0147
XFMR Power (W)	0.002	0.002
Driver Only Power (W)	0.522	0.558
Total Supply Power Consumed (W)	**0.6**	**0.684**

Figure 23. AD8393 System Measurements

7. Conclusions

Clearly much progress has been made with Class AB amplifiers in reducing the TSPC in ADSL applications. However, this power reduction has asymptotically reached it's maximum efficiency and further reduction will be limitied. Also, with the emerging ADSL2/2+ standards along with the desire to broadcast video on demand over phone lines, the TSPC of Class AB amplifier systems will increase beyond today's levels back close to 1W. With these new requirements and the nature of the DMT waveform used in ADSL, the environment for Class G and Class H/H-like amplifiers is favorable for power reduction beyond today's Class AB amplifiers even with the emerging standards and applications. The work presented here not only provides a theoretical basis for these claims, but also basic circuit concepts to help understand how power reduction is possible with Class G and Class H/K-like amplifiers.

Acknowledgements

There are several contributors to the work above this author wants to acknowledge. First, to Scott Wurcer for the initial Class H-like architecture and all of the theory presented comparing the efficiencies of the different types of amplifier systems. Second, to Bob Day for analog and system verification of the AD8393 Class H-like amplifier. Additionally, the author wants to acknowledge contributions from the Wilmington Fab/Prototype Assembly, members of Analog Devices Communications Division, Pete Commerford for evaluation board designs, and Steve Robichaud for the many layout variants. Special thanks goes to Lew Counts for numerous technical discussions during the design phase of this work and his continued support of the efforts from all individuals working on these systems.

References:

[1] M. Corsi, R. Hester, K. Maclean, M. Agah, J. Quarfoot, C. Kozok, N. Gibson, T. Hagen, “An ADSL Central Office Analog Front-End Integrating Actively-Terminated Line Driver, Receiver and Filters”, Paper 19.4, ISSCC 2001.

[2] F. Sabouri, R. Shariatdoust,”A 740mW ADSL Line Driver for Central Office with 75dB MTPR”, Paper 19.2, ISSCC 2002.

[3] R. Benton, R. Apfel, B. Webb, J. Wenske, W. Schopfer, F Thiel,”A High-Voltage Line Driver (HVLDR) for Combined Voice and Data Services”, Paper 19.3, ISSCC 2001.

[4] D. Whitney, “Circuit Adapts Differential Input to Drive Coax”, EDN, May 8th, 1997.

[5] Tripath Data Sheet, TLD401x, 2001.

[6] K. Macclean, et. Al., “A 610mW Zero-Overhead Class G Full-Rate ADSL CO Driver”, Paper 23.4, ISSCC 2003.

[7] J. Pierdomenico, S. Wurcer, B. Day, “A 684mW Adaptive Supply Full Rate CO Driver”, JSSCC, December 2002, pp. 1831-1838.

[8] S-S Lee, et. Al, “A 700mW CMOS Line Driver for ADSL Central Office Applications”, Paper 23.5, ISSCC2003.

[9] SoundCraftsman MA5002, 1978.

[10] Philips TDA8574 & TDA1562Q, 1998.

[11] J. Quarfoot, ”Managing ADSL Signals and Contending with Noise”, CDS Magazine, December, 1998.

The USB 2.0 Physical Layer: Standard and Implementation

Gerrit W. den Besten
Philips Research Laboratories Eindhoven
Prof. Holstlaan 4, 5656 AA Eindhoven
The Netherlands
Tel/Fax: +31 40 27 42653/44657
e-mail: gerrit.den.besten@philips.com

Abstract

Realization of the electrical-physical layer for the USB 2.0 standard requires sophisticated mixed-signal electronics. The combination of the new 480 Mb/s high-speed mode together with backward compatibility with USB 1.1 is technically challenging. In this paper the essential parts of the USB protocol are described and important system trade-offs are discussed. Furthermore, the structure and operation of the PHY are explained and implementations of some critical building blocks are shown. The use of integrated calibrated resistors in order to minimize the number of external components is described. Finally, silicon realizations are shown and measured results are presented.

1. Introduction

The Universal Serial Bus (USB) is a widely accepted interface for PC peripherals today. Ease-of-use and the possibility of low-cost implementation have contributed significantly to its success. The low-speed (LS) and full-speed (FS) modes, at 1.5 and 12 Mb/s respectively, were defined in the USB 1.0 standard in January 1996 [1] and in its revision, USB 1.1, in 1998 [2]. A consortium of companies composed of Intel, Microsoft, Compaq, Philips, HP, Lucent, and NEC started working on the extension of the bandwidth in early 1999. This resulted in the addition of a new 480 Mb/s high-speed (HS) mode, which has been included in the USB 2.0 standard, released in April 2000 [3].

The structure of a general USB network is hierarchical, with one host controller (usually a PC) at the top tier and up to 127 device functions at the endpoint nodes. The network is constructed using cables, each of which can be up to 5 m long, and hubs, which route and distribute the signals between devices and the host (cf. Figure 1) The network is allowed to have up to 6 tiers,

A. van Roermund et al. (eds.), Analog Circuit Design, 359-378.
© 2003 *Kluwer Academic Publishers. Printed in the Netherlands.*

corresponding to a host with 5 cascaded hubs, which implies a maximum of 6 cable connections between the host and any device. Hot plug-and-play, foolproof operation, and high fault tolerance are important features of USB.

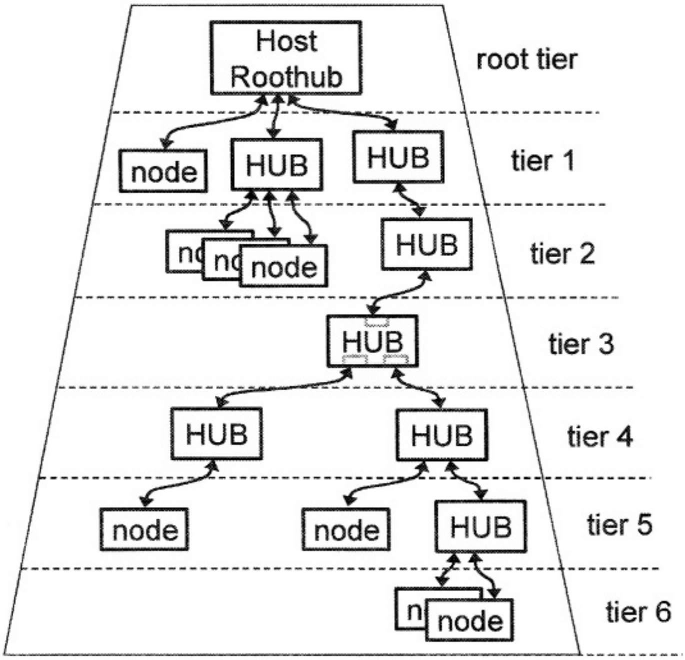


Figure 1: USB topology

In a classic USB network all communication is initiated by the host, which polls all devices. During initialization or reset of any connection the mode of operation is negotiated, after which packet based communication becomes available at the highest possible speed.

USB establishes on FS/LS connections a 1 ms time base, called a frame, and on high-speed connections a 125 µs time base, called a micro-frame. Frames and micro-frames are indicated by Start-of-Frame packets (SOF, micro-SOF) and may contain several transactions.

The USB 2.0 standard is backward compatible with USB 1.1, which means that devices from both standards can be used in the same network. However, the HS mode of any device can only be exploited if everything from the host down to that device is HS capable. In order to prevent lower speed traffic holding up the high-speed bus traffic in the upper tiers, speed switching is applied. The first HS capable hub in the chain converts the LS and FS packets to HS transactions for further communication with the host. Issues of this type are handled at the protocol/hub layer.

The recent development of USB On-The-Go (OTG) enables the possibility of making a single point-to-point connection between two devices. OTG targets cases where no PC is available, like for example the connection of a digital camera to a printer. An OTG peripheral can have the following enhancements: limited host capability to communicate with other selected USB peripherals, a small USB connector to fit the mobile form factor, and low power features to preserve battery life.

The remainder of this article describes the electrical/physical layer of the USB 2.0 standard. This includes channel characteristics, transmission modes and interoperability, line-coding and signaling, packet formats, phase alignment, and data transmission, reception, and repetition.

The physical layer interface is the connecting module between the cable and the inner protocol logic for every port of a device. This is illustrated for the HUB at tier level 3 in figure 1.

A block diagram of the USB physical layer interface is shown in Figure 2. This sophisticated mixed-signal module contains output drivers, receivers, level detectors, terminations, high-speed logic (for CODEC, SERDES, FIFO and data recovery), and reference circuits, such as a band gap, power-on-reset, crystal oscillator, and PLLs. The blocks in the dashed box are also needed for USB 1.1, although control functionality has increased. However, most of the circuitry is inside the dotted box, which is the part required for high-speed operation. Additionally, the resistors for line termination and protocol signaling, which were previously realized externally, are now integrated inside the IC.

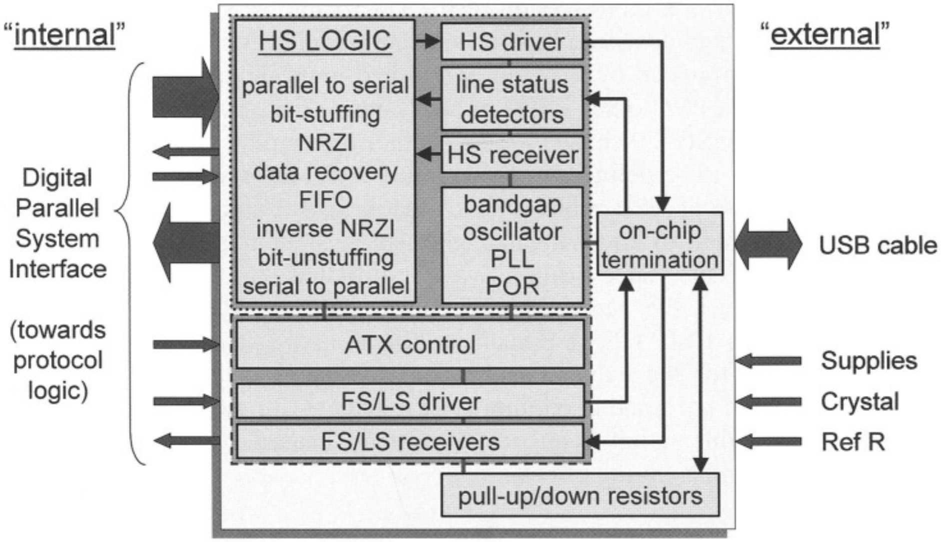

"internal"
"external"
HS LOGIC
parallel to serial
bit-stuffing
NRZI
data recovery
FIFO
inverse NRZI
bit-unstuffing
serial to parallel
HS driver
line status detectors
HS receiver
bandgap
oscillator
PLL
POR
on-chip termination
ATX control
FS/LS driver
FS/LS receivers
pull-up/down resistors
Digital Parallel System Interface
(towards protocol logic)
USB cable
Supplies
Crystal
Ref R

Figure 2: USB Transceiver block diagram

2. The transmission channel

A connection between any two devices consists of a cable, two pairs of connectors (A/B - male/female) and some PCB tracks at both ends, which carry signals to the ICs. For the relatively low bit rates of USB 1.1 (1.5/12 Mb/s) the channel requirements are pretty loose compared to what is required in USB 2.0 for HS at 480 Mb/s.

For USB 2.0, PCB requirements are given by a TDR specification, which specifies nominal characteristic impedances, permitted level of discontinuities (parasitic components), and PCB track delays.

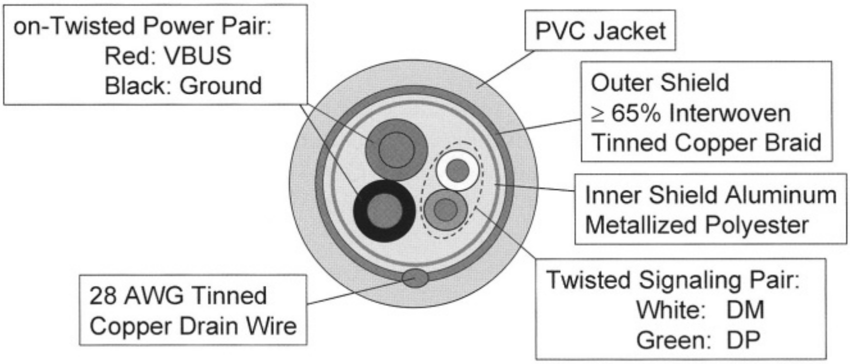

on-Twisted Power Pair:
Red: VBUS
Black: Ground
PVC Jacket
Outer Shield
≥ 65% Interwoven
Tinned Copper Braid
Inner Shield Aluminum
Metallized Polyester
28 AWG Tinned
Copper Drain Wire
Twisted Signaling Pair:
White: DM
Green: DP

Figure 3: Cross-section of a USB cable

Figure 3 shows a cross-section of the USB cable as specified in the 1.1 standard. It consists of a twisted pair for the data signal, a supply pair, and some filler material, surrounded by a shield of foil and braid including a drain wire, and finished with a PVC outer isolation [1, 2]. One of the issues that had to be addressed for the USB 2.0 standard was whether or not new types of cable and connectors needed to be defined for the HS data rate. However, characterization of the installed base of USB cables showed that USB 1.1 compliant cables were actually good enough to allow much higher-speed signaling. Therefore, slight modifications of the specification were sufficient to allow HS mode communication over the same type of cables, thus ensuring full backward compatibility with USB 1.1. One such modification was the extension of the frequency range, for the cable transmission loss specification, to 400 MHz. Figure 4 shows the specified maximum loss, together with measured results, for a typical 5 m cable. High-frequency losses are caused by the skin effect, dielectric losses, and coupling effects.

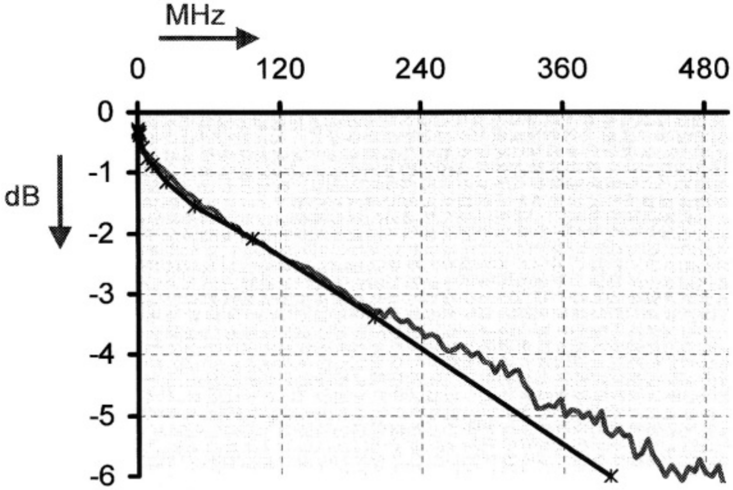


Figure 4: Specified and measured losses of a 5 m USB cable

Furthermore, maximum skew and delay are more tightly specified at 100 ps and 5.2 ns/m respectively. A common-mode characteristic impedance specification of 30 Ω ± 30% has been added, while the differential characteristic impedance specification is maintained at 90 Ω ± 15%.

Given the specified characteristics of both the cable and the total channel, requirements for USB transceiver modules are derived and defined by means of eye templates at the connectors for receive and transmit, including test conditions. In this way reliable data transfer is guaranteed and interoperability between PHYs of different vendors is enforced.

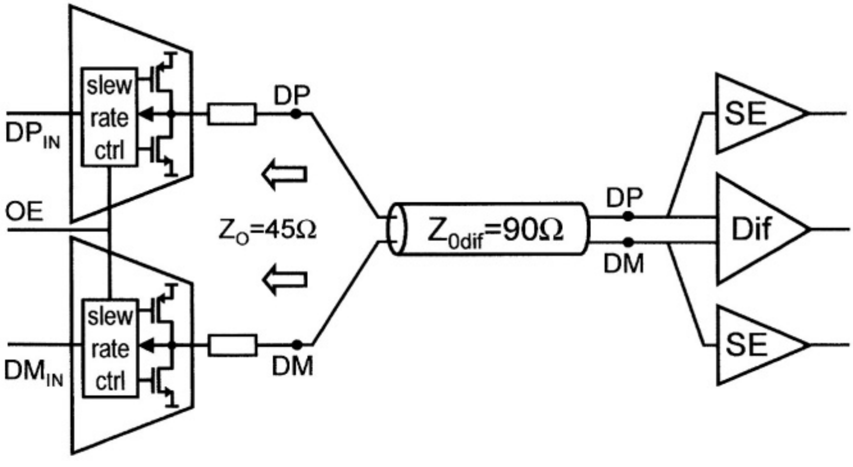


Figure 5: LS/FS driver and receiver scheme

3. LS and FS mode signaling

The signaling level for LS and FS is 3.3 V, just as in USB 1.1. The cable is driven by a pair of slew-rate controlled drivers, with an output impedance specified at 45 Ω ± 10% for USB 2.0 (Figure 5). The source termination scheme implies full-swing signals on the bus without DC power consumption. However, this comes at the cost of full signal reflection at the receiver side. The slew-rate control reduces EMI and enhances symmetry of data crossovers. Three of the four possible states of the pseudo-differential driver are used: differential '1' and '0' states for data packets, and single-ended-zero (SE0) for protocol handling. The condition where both outputs are high (SE1) is a forbidden state (Table 1).

The analog front-end of the receiver consists of a differential comparator-amplifier to recover the data and two single-ended detectors to detect the state of the bus. Hubs in LS and FS mode amplify the signal and re-transmit it, thereby largely removing reflections and glitches, but do not perform timing recovery.

		Modes		
States	DP/DM	HS	FS	LS
Differential '1'	H/L	J	J (idle)	K
Differential '0'	L/H	K	K	J (idle)
SE0	L/L	idle	EOP/reset/disconnect	
SE1	H/H	illegal	illegal	illegal

Table 1: Overview of line states

4. HS mode signaling

In HS mode, low-swing (400 mV) differential signaling is applied. In this case, the bus on both sides is terminated to ground with 45 Ω ± 10% resistors, giving the correct differential impedance and a reasonable common-mode impedance match (Figure 6). This results in high robustness, while keeping power consumption at an acceptable level. Termination resistors can be reused because a USB 1.1 driver, which drives SE0, gives the right terminations for HS mode.

During data transmission a differential pair current driver switches a constant current between the two outputs. This current flows into the termination resistors, thereby generating a differential voltage swing. In the absence of data packets, the driver current can be drained to ground via a third switch path, or the current source may be switched off. In practice a combination of both of these measures is employed to reduce power consumption and relax the on/off switch-time requirements of the source.

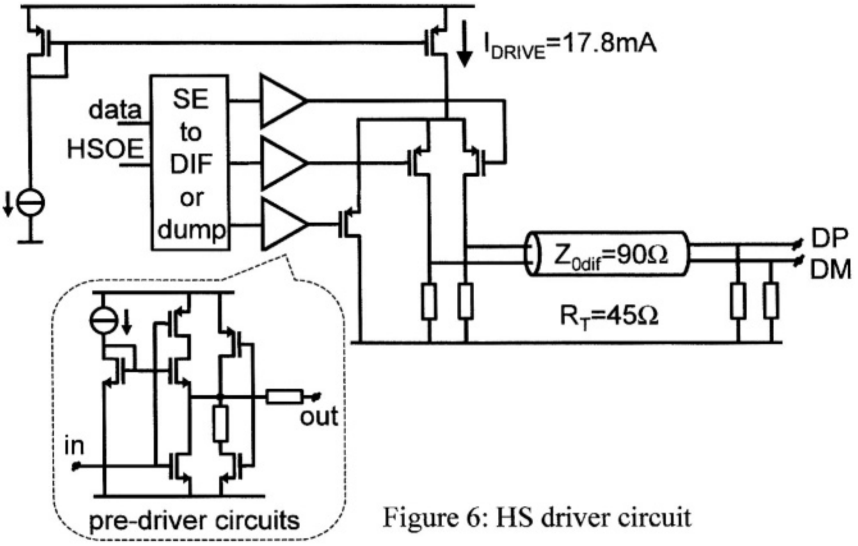


Figure 6: HS driver circuit

Limited-swing pre-driver circuits reduce capacitive throughput. Their low-crossing point for transitions minimizes the ripple at the common tail node. Pre-driver outputs are pulled to the positive supply during other modes in order to prevent large leakage currents via the PMOS output devices.

Switching off the output current brings the bus into HS Idle state, which equals SE0 for LS and FS. The resulting common mode steps cause no problems if sufficient common mode input range is preserved in the differential receiver circuits. The reception of signals is fully differential and a differential input amplifier stage delivers the input signal to the data sampling and recovery unit.

Figure 7 shows measured differential HS signals at the near-end and far-end connector of a 5 m cable at the beginning of a packet. This figure clearly illustrates the delay and losses of the channel.

In order to distinguish data signals on the bus from noise and disturbance, a squelch detector is included, which measures when the differential receiver input signal amplitude exceeds a certain threshold between 100-150 mV. The detector must indicate squelch when the magnitude of the differential voltage envelope is $\leq$ 100 mV, and must not indicate squelch if the amplitude of differential voltage envelope is $\geq$ 150 mV. The implementation of the squelch circuit is shown in Figure 8. The circuit compares the level-shifted differential input signal with the reference voltage levels across resistors R1 and R2. These reference levels are derived from a band gap reference circuit. If the input amplitude exceeds the reference level in either direction, one of the comparators switches and node ‘s’ is forced high. The current limited discharge of this node provides a filtering function, which prevents the squelch circuit output reacting to differential data crossovers.

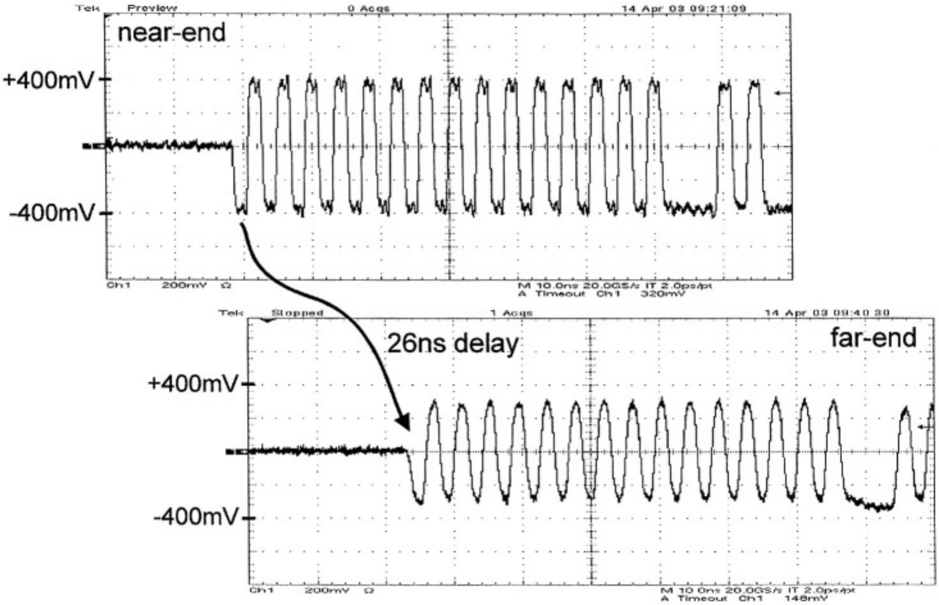


Figure 7: Differential HS signals at the near-end and far-end of a 5 m cable

In HS mode, due to much higher frequencies, signal dispersion in the channel is too severe to repeat the signal in a hub by amplification only, as in LS/FS modes. For this reason, every hub must recover incoming HS data packets and re-time the data to a local clock before repeating them. For these hub repeaters there is, besides the squelch detector, an additional check on the synchronization sequence (SYNC) at the beginning of a packet in order to reduce the probability of noise propagation in the system.

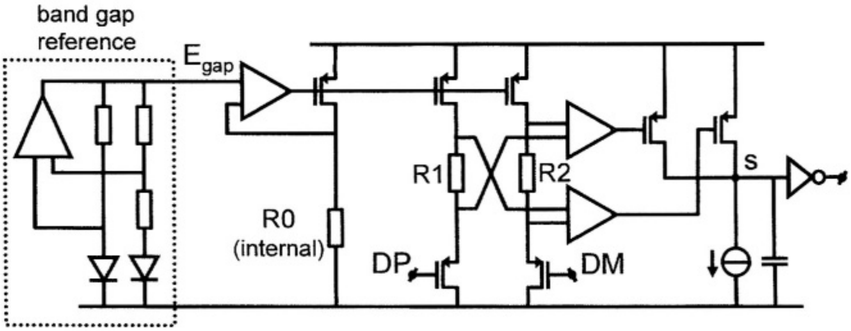


Figure 8: Squelch detection circuit

5. Mode detection

Apart from data signaling, there are mechanisms implemented to determine the mode of operation between two PHYs in order to exploit the highest possible bit rate. Figure 9 shows the required configuration around the cable for this 'mode signaling'. This includes the previously discussed LS/FS and HS drivers and some additional resistors. Note that the LS/FS drivers, driving SE0, provide a correct termination for the HS current driver.

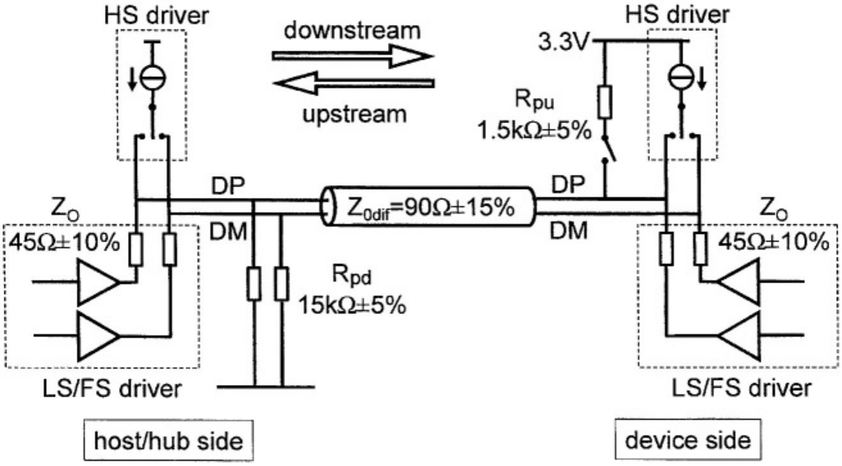


Figure 9: Mode detection configuration

In accordance with the structure of the bus, 'mode signaling' is always top-down. At the host side of each connection there are two 15 kΩ pull-down resistors Rpd, which generate a soft low on the bus (soft SE0) when nothing else is driving the bus actively high. Furthermore, the termination resistors and several drivers for each mode can all be disabled separately. Finally, in upstream facing ports, there is a 1.5 kΩ pull-up resistor Rpu connected to either DM or DP, for LS and FS/HS devices respectively, which must be switchable for HS capable devices. This pull-up resistor creates a differential '0' on the bus for LS and a differential '1' for FS/HS devices, assuming no driver is active and HS terminations are disabled.

At the instant of initialization of a connection a downstream facing port is in soft SE0 state due to the 15 kΩ pull-down resistors. If a LS device is connected, its pull-up resistor on DM generates a differential '0' (LS-J), which is detected by the legacy level detectors on the host side, and the connection enters LS mode. A long SE0 (10 ms) is transmitted downstream to reset the connected device. Before the HS mode was specified, the procedure for FS was similar except that, in that case, a DP pull-up generates a differential '1' (FS-J) on the bus. However, in the USB 2.0 standard this procedure has been extended to allow an additional HS mode while maintaining backward compatibility with USB 1.1 devices.

A speed negotiation protocol, known as 'chirping', has been introduced to allow a USB 2.0 hub/host to detect USB 2.0 device connections. If a HS capable, downstream facing port detects a differential '1' on the bus, it sends an SE0 and starts a timer. If a HS-capable device is connected to the downstream side, it should respond with a Chirp K on its upstream port within 7 ms for a fixed amount of time. Chirp K means that the HS current driver of that device drives a K state on the lines without enabling its HS terminations. This generates a K state on the bus, with approximately 800 mV amplitude. This level is too low to be detected by USB 1.1 level detectors, but is easily recognized by the HS capable, downstream facing port. Note that, without driving a Chirp K, the bus would show a 'tiny J' because of the FS pull-up resistor. After receiving the Chirp K, the downstream facing port switches to HS mode because it knows that the connected device is HS capable. After the Chirp K ends, it responds with a fixed number of slowly alternating Chirp J and K states. This tells the connected device that the downstream facing port is also HS capable. When the device recognizes this sequence it switches to HS mode, which completes the chirp protocol.

The chirp will fail if one side of the link is not HS capable, because either the Chirp K is not transmitted, or the downstream port will neither detect it nor respond to it. In this case both sides will enter FS mode. The situation where both sides are not HS capable corresponds to the conventional USB 1.1 case. The chirp procedure is depicted in Figure 10.

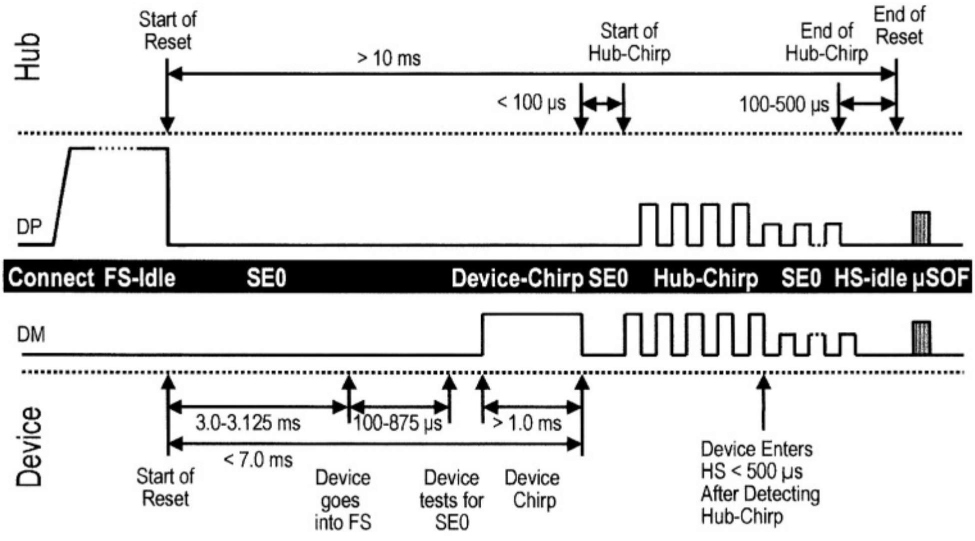

Hub
Start of Reset
> 10 ms
Start of Hub-Chirp
End of Hub-Chirp
End of Reset
< 100 µs
100-500 µs
DP
Connect FS-Idle SE0 Device-Chirp SE0 Hub-Chirp SE0 HS-idle µSOF
DM
Device
3.0-3.125 ms
100-875 µs
> 1.0 ms
< 7.0 ms
Start of Reset
Device goes into FS
Device tests for SE0
Device Chirp
Device Enters HS < 500 µs After Detecting Hub-Chirp

Figure 10: Chirp or speed detection protocol

6. Data encoding

In order to balance data and generate sufficient data transitions on the bus, NRZI encoding with bit stuffing is employed. NRZI means that, for each zero, the line state is inverted, whereas for ones the line state remains unchanged. In order to limit the maximum run length in case of long strings of ones, a zero is 'stuffed' into the stream after every 6 consecutive ones, thus forcing a transition of the line state. This operation, known as bit stuffing, is performed even when the 7th bit is a zero, in order to avoid ambiguity at the receiver. This method implies that there is at least one signal transition every 7 bit periods.

This encoding method trades a small data rate penalty (< 1% for random data), for improved performance. The drawback is that this makes the effective throughput slightly data dependent, which implies that the speed of the PHY's parallel I/O data busses at the system side, need to be controlled to match the bus transfer rate. This is implemented by holding these busses for an extra clock cycle, when the PHY has encountered so many bit stuffs that it is still processing the previous parallel word.

7. Packets

A packet (Figure 11) is composed of a Synchronization (SYNC) sequence, a packet field, and an End-of-Packet (EOP) sequence. The packet field is made up of a Packet Identifier (PID), a field (DATA) containing control bits, address information, or user data depending on the packet type, and a Cyclic Redundancy Check (CRC) code, for error detection.

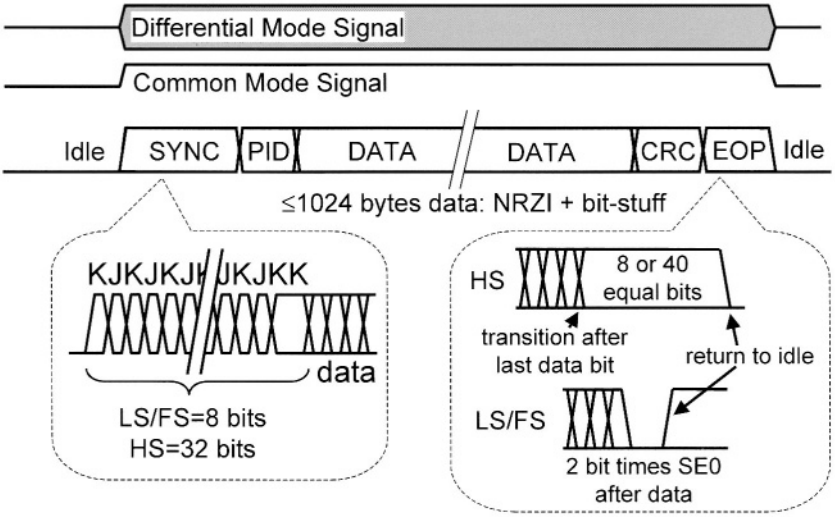

Differential Mode Signal
Common Mode Signal
Idle
SYNC
PID
DATA
DATA
CRC
EOP
Idle
≤1024 bytes data: NRZI + bit-stuff
KJKJKJK JKJKK
data
LS/FS=8 bits
HS=32 bits
HS
8 or 40 equal bits
transition after last data bit
return to idle
LS/FS
2 bit times SE0 after data

Figure 11: Packet format

Although the structure of packets is similar for different modes, the sequences for SYNC and EOP are different. In LF/FS mode the SYNC consists of 8 bits, KJKJKJKK, whereas in HS mode a SYNC consists of a total of 32 bits composed of alternating Ks and Js and ending with two Ks.

The process of squelch detection, clock timing, and data recovery in HS mode may result in a loss of SYNC bits in repeaters. For this reason, a longer HS SYNC has been defined to ensure that the SYNC is sufficiently long at all tier levels to detect packets and synchronize timing before the data content starts. The loss of HS SYNC bits is known as pre-dribble.

Each hub is permitted to sacrifice up to 4 bits of the SYNC for repeating, but it is not allowed to corrupt the pattern. This implies that for 5 hubs in cascade at least 12 SYNC bits will be left at the endpoints.

A HS packet can contain up to 1024 bytes of un-coded data, which makes the worst-case packet length about 10k bits. A HS EOP at the end of a packet is realized with a string of 8 equal consecutive bits on the bus (40 for the long EOP used in micro-SOF), which violates the bit-stuff encoding rule. This event is not allowed inside data packets so that, when a bit-stuff error is detected, it is assumed that the packet has ended. Furthermore, after the EOP, the bus returns to HS idle, which is detected by the squelch detector. In the same way as SYNC bits are sacrificed at the beginning of packets in repeaters, noise bits may be added to the end of the packet, which is called post-dribble. HS hubs are allowed to add up to 4 bits of post-dribble, but must not corrupt the EOP.

In LS and FS modes an EOP consists of driving SE0 for 2 bit times, before the bus returns to the idle (J) state, which is signaled by the level-detectors. Pre- and post-dribble will not occur because data recovery is not included in hubs for these modes.

8. Clock domains

Each device in the USB network has its own reference clock provided by a local crystal oscillator. Tolerance on the accuracy of the frequency is specified at ± 500 ppm. This means that for HS mode the accumulated timing misalignment over one packet can be as high as 10 bits. Because hubs, which recover and re-time the data, must be able to repeat packets without corrupting them, an elastic buffer is needed to avoid underflow or overflow situations. Therefore, a FIFO buffer of sufficient size is partially pre-loaded with data before the repeat output stream is started. This FIFO has cyclic, gray-coded, read and write address pointers for glitch free simultaneous asynchronous read and write functionality.

9. Data recovery unit

A block diagram of the data recovery unit is shown in Figure 12. A PLL generates multi-phase 120 MHz clock signals from an external 12 MHz crystal reference clock. The pre-amplified data signal is oversampled with samplers driven by the multi-phase clock signals. The sampler outputs are re-timed to a single 120 MHz clock phase and stored in a register. Edge information is extracted and delivered to the edge tracking and phase selection unit which tracks sliding phase due to frequency mismatch and, at the same time, filters out short term variations. The sampler phase, which is deemed to contain the samples closest to the middle of the data eye, is selected and the corresponding 4 samples are stored [4, 5].

A consequence of the sliding phase is that the selected sampler phase will drift and, after some time, an under-run or over-run situation will occur where either a bit will be lost or included twice. This problem has been solved by storing 5 or 3 bits instead of 4 bits in those cases. These situations can be easily detected by monitoring the phase jumps. The resulting 3 to 5 bit data words, synchronized to the local clock, are sent to the FIFO, together with under- and over-run status bits.

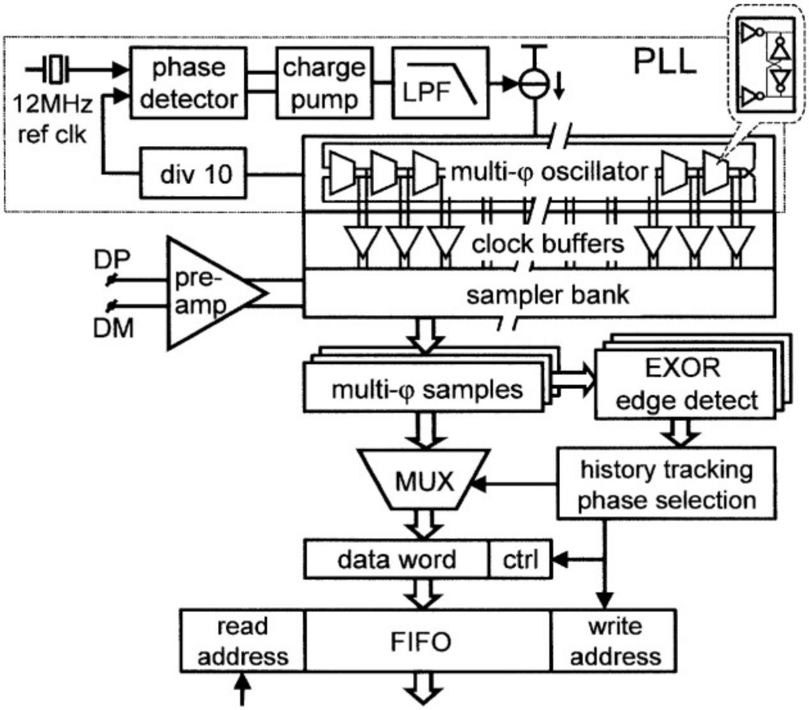

PLL
12MHz
ref clk
phase detector
charge pump
LPF
div 10
multi-φ oscillator
clock buffers
DP
DM
pre-amp
sampler bank
multi-φ samples
EXOR edge detect
MUX
history tracking phase selection
data word
ctrl
read address
FIFO
write address

Figure 12: Data recovery unit

An important requirement for data recovery in hub repeaters is the amount of latency added to the data stream. In order to minimize the total bus turn-around time, the repeater must have a latency of less than 36 HS bit times plus 4 ns, measured between input and output connectors. As the pre-loading of the FIFO accounts for a significant portion of this budget, signal processing for timing recovery is performed in parallel to the data flow itself.

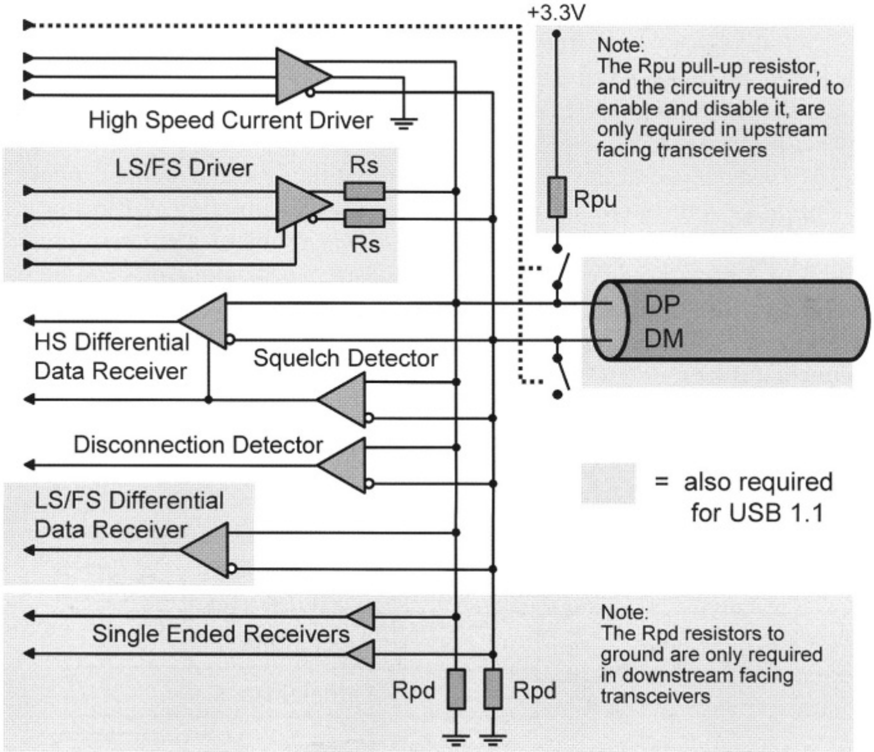

+3.3V
Note:
The Rpu pull-up resistor, and the circuitry required to enable and disable it, are only required in upstream facing transceivers
High Speed Current Driver
LS/FS Driver
Rs
Rs
Rpu
DP
DM
HS Differential Data Receiver
Squelch Detector
Disconnection Detector
= also required for USB 1.1
LS/FS Differential Data Receiver
Single Ended Receivers
Note:
The Rpd resistors to ground are only required in downstream facing transceivers
Rpd
Rpd

Figure 13: I/O circuitry for HS capable devices

10. Disconnect

Unplugging a cable, or any other line-break event, causes a loss of connection between two PHYs. This disconnect situation must be detected by the remainder of the network in order to prevent useless communication on this port. For both LS and FS this event is seen on the bus because the pull-up of the disconnected device will fail to generate a J state (differential '0' for LS and differential '1' for FS) and an SE0 will be seen on the bus during Idle status between packets.

In the case of disconnection in HS mode, the bus status does not change during Idle mode because nothing is driving the bus anyway. However, the terminations at the downstream side of the link are disconnected now. This means that the amplitude of HS signals driven on the lines doubles from 400 mV to 800 mV. Because the lines are not properly terminated anymore the actual signal may behave in a very strange way for regular data content.

A micro-SOF (Start-of-Frame, which schedules the bus time) with a long EOP, consisting of 40 equal bits, has been defined to facilitate disconnect detection in HS mode. This long period of stable output guarantees sufficient signal settling, taking into account the propagation delay of the cable. A disconnect detection circuit on all HS capable downstream facing ports, which operates in a similar way to the squelch circuit, determines when the signal level exceeds a threshold within the range 525 mV - 625 mV. Figure 13 shows the arrangement of all the required circuitry around the I/O pins for HS capable devices.

11. Reference voltages and integrated resistors

USB 2.0 signaling requires both accurate voltage levels and precision resistors. An integrated band gap reference circuit provides an accurate absolute reference voltage. By means of V/I and I/V conversions using internal resistors of the same type, scaled values of the band gap voltage can be accurately derived. This principle has, for example, been applied in the squelch detector circuit described before.

Line termination, pull-up, and pull-down resistors have, traditionally, been implemented off-chip because of the required accuracy. However, in this implementation most resistors are integrated. Using one precision external reference resistor, currents with accurate absolute values can be generated. With these currents, digitally controllable internal resistors can be tuned to a preferred resistance level with a simple calibration loop. The precision is mainly limited by characteristics of the controllable resistor, such as tuning step size and linearity. The tuning concept is depicted in Figure 14.

The known current derived from the external resistor is forced through a replica of the internal resistor/MOS structure that is used for the terminations. Comparison of the resulting voltage Vi with Egap gives the digital code that is used to select the correct set of similar, but scaled, parallel branches. As only PVT variations have to be calibrated, the greater part of the resistance is kept fixed, which reduces the influence of the branch switching on the total resistance. The calibration is performed during reset.

In this way, it has been possible to integrate termination and pull-up resistors, which eliminates a lot of external components. Furthermore, it saves 1C pins, since both sides of the termination resistors are actively driven, thus necessitating two pins per resistor for external solutions. The accuracy of the resistors after calibration is ± 10% for the 45 Ω terminations and ± 5% for the pull-up and the pull-down resistors, which conforms to the USB specifications.

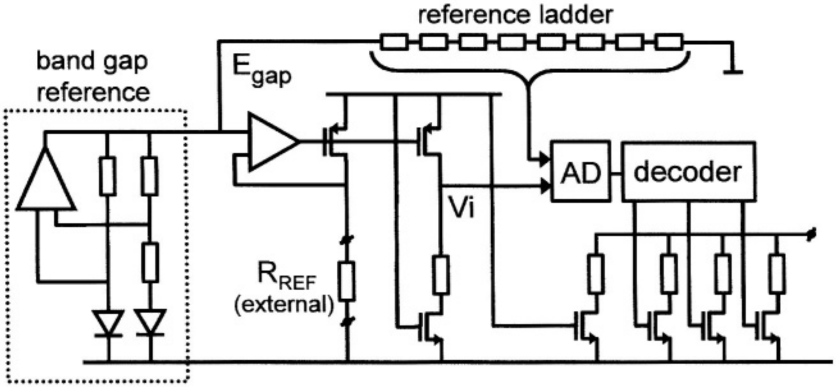


Figure 14: On-chip resistor calibration scheme

12. The parallel system interface of the PHY

The system interface of the PHY contains a dual uni-directional or single bi-directional parallel data bus. This bus may be implemented either 16, 8, or 4 bits wide, running at 30, 60, or 120 Mb/s/pin respectively, depending on pin count and clock speed preferences. A handshake procedure, using 'Valid' and 'ready' signals, is used to maintain alignment between data on the parallel interface and data on the USB bus, thus compensating for the slight difference of effective data rates due to bit-stuffing. In addition to this data bus, the interface also includes control signals, such as enable signals for several functions within the PHY, and detector output signals, which are needed by the higher-level protocol. Recently, the UTMI+ standard has been released in an effort to standardize this parallel interface format [6, 7, 8]. However, it is beyond the scope of this paper to explain this interface in further detail.

13. Process technology issues

The first USB 2.0 transceiver was implemented in a 0.35 µm CMOS process [9]. Because 480 MHz was rather difficult in this technology, parallelism and pipeline tricks were applied to implement data recovery and processing. The problem with smaller feature size processes is that the LS/FS signaling level is 3.3 V, which requires almost 5 V tolerance when reflection effects are included. For this reason, thick gate oxide devices are applied in circuits connected to the I/O pins, while internal circuitry with normal gate oxide, operating at standard process supply voltage, profits from increased speed and power reduction. In this way the design has been successfully ported to 0.18 µm and 0.13 µm CMOS processes.

14. Testability

In order to make this mixed-signal block testable on mass production testers the block has a CTAG.AMS shell [10]. The CTAG.AMS test strategy allows both for scan tests of the digital part of the transceiver and analog functional tests. Special analog modes enable external monitoring of internal transceiver signals. At-speed loop-back tests are implemented and measures have been taken inside the analog part of the transceiver to enable DC-tests on essential USB parameters. By means of this strategy a high test coverage is achieved in the transceiver.

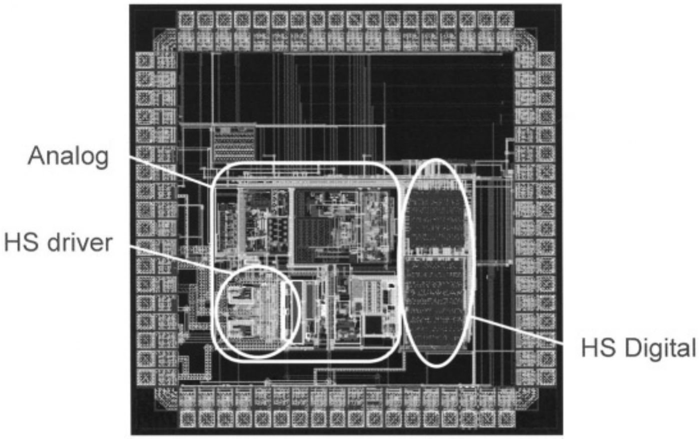


Figure 15: Test silicon transceiver layout

15. Results

The transceiver has been implemented in several technologies. Figure 15 shows a layout of test silicon. The silicon die area is about $0.8\ \text{mm}^2$. The PHY is measured using a PHY test PCB plugged into a generic FPGA board to generate all possible data and control signals to test the circuitry. Figure 16 shows a photograph of a test board. A product evaluation board, which includes a transceiver chip, is shown at the bottom left-hand side of the picture for comparison of sizes.

Figure 17A shows the HS transmit eye, with template, measured at the near-end connector while Figure 17B shows the received HS eye patterns at the far-end connector. Both measurements were performed using a typical 5 m cable connection.

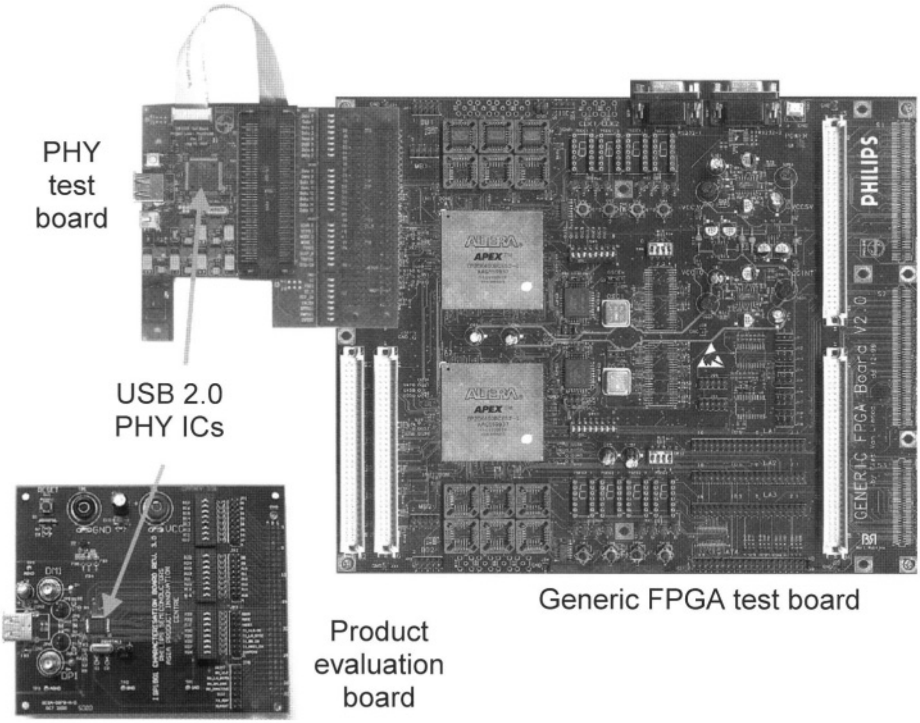

PHY test board
USB 2.0 PHY ICs
Generic FPGA test board
Product evaluation board
ALTERA
APEX
PHILIPS
GENERIC FPGA Board V2.0

Figure 16: Measurement boards

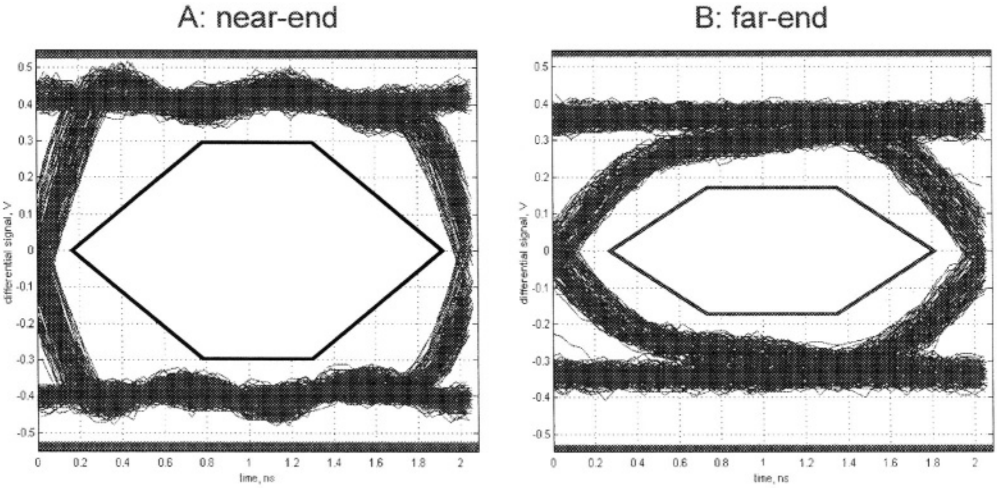
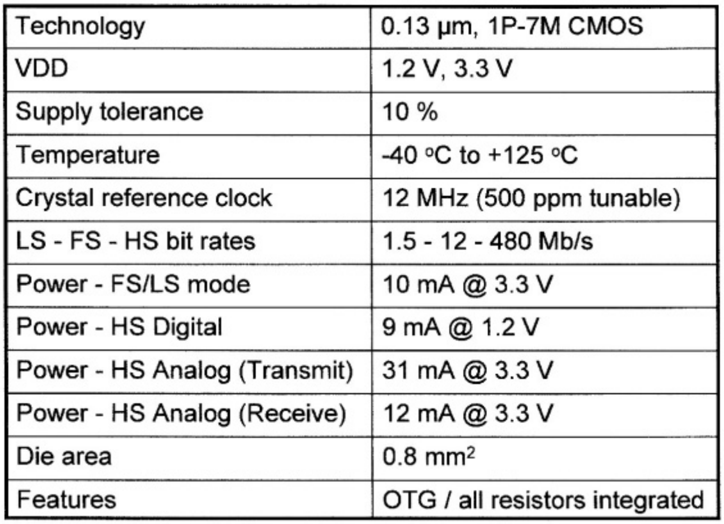


Figure 17: Eye diagrams at near-end and far-end of a 5 m cable, including the specified eye templates

The power consumption of the PHY, including output drivers, varies between 30-130 mW depending on the (active) state and mode of operation. The HS current output driver consumes about 60 mW, which pretty much dominates the total power budget. In power-down mode (Suspend) there is hardly any power consumed apart from about 200 µA in the pull-up and pull-down resistors. Table 2 summarizes the performance of a 0.13 µm CMOS PHY example.



Technology	0.13 µm, 1P-7M CMOS
VDD	1.2 V, 3.3 V
Supply tolerance	10 %
Temperature	-40 °C to +125 °C
Crystal reference clock	12 MHz (500 ppm tunable)
LS - FS - HS bit rates	1.5 - 12 - 480 Mb/s
Power - FS/LS mode	10 mA @ 3.3 V
Power - HS Digital	9 mA @ 1.2 V
Power - HS Analog (Transmit)	31 mA @ 3.3 V
Power - HS Analog (Receive)	12 mA @ 3.3 V
Die area	0.8 mm^2
Features	OTG / all resistors integrated

Table 2: Performance of transceiver example

16. Acknowledgements

I would like to thank all my colleagues who contributed to circuit design, measurements, discussions, this paper, and/or the USB 2.0 standardization process. In particular I'd like to mention Roland de Jonge, Chung Wing Yan, Conor O'Donoghue, Farran MacKay, Jerome Tjia, Jerry O'Mahony, Conar Dunne, Marq Kole, and Roland van Wegberg.

17. References

[1] USB 1.0 standard, January 15, 1996.

[2] USB 1.1 standard, September 23, 1998. Available at: *www.usb.org/developers/docs/usbspec.zip*

[3] USB 2.0 standard, April 27, 2000. Available at: *www.usb.org/developers/docs/usb20.zip*

[4] K. Lee *et al.,* "A CMOS serial link for fully-duplexed data communication," *IEEE J. Solid-State Circuits,* Vol. 30, No. 4, pp. 353-364, April 1995.

[5] C-K. Yang, R. Farjad-Rad, and M.A. Horowitz, "A 0.5μm CMOS 4.0-Gbit/s serial link transceiver with data recovery using oversampling," *IEEE J. Solid-State Circuits,* Vol. 33, No. 5, pp. 713-722, May 1998.

[6] Philips transceiver interface specification. Available at: *www.semiconductors.philips.com/acrobat/datasheets/ISP1501-02.pdf*

[7] UTMI specification, version 1.05, March 2001. Available at: *www.intel.com/technology/usb/spec.htm*

[8] UTMI+ specification, Available at: *www.semiconductors.philips.com/buses/usb/products/utmiplus/index.html*

[9] W.L. Chui and G.W. den Besten, "USB 2.0 silicon building blocks," *USB Developers Conference,* October 1999, San Diego.

[10] C. Feige and G. Seuren, "Extending core test methodology to the analog/mixed-signal domain," *European Test Workshop 2001,* Stockholm.

BACKPLANE TRANSCEIVERS

Kimo Tam, William Ellersick, Richard Soenneker
Analog Devices
Wilmington, Massachusetts, USA

Abstract

High-speed serial digital backplane transceivers must equalize broadband channel loss at frequencies up to several GHz. Power dissipation must be minimized to allow for integration of many transceivers on a single IC. This paper will include a system level overview of the electrical backplanes typical in state-of-the-art network equipment. The signal integrity considerations applicable to this channel will be presented. Mixed-signal circuit approaches to implement highly integrated backplane transceivers that are interoperable with conventional binary signaling will be discussed. The tradeoffs of multi-level signaling for future applications will also be covered.

1. Introduction

The capacity requirements for high performance backplanes for networking and computing applications, measured both in terms of aggregate bandwidth as well as bandwidth-distance product per link, continue to rise. In the near future, several hundred Gbits/s will be concentrated into compact systems having backplanes on the order of $0.5 \times 0.5\ m^2$, with 20 or more layers. To support these system bandwidths, the speed of the backplane links must grow. Parallel, multi-drop busses clocked at a few hundred MHz will be replaced by multi-Gb/s, controlled-impedance, point-to-point serial links routed through a central or distributed switch. A variety of competing standards exist for serial backplane communications; these include SPI-5, XAUI and Infiniband. These standards commonly bundle multiple 2.5 Gb/s or 3.125 Gb/s links for higher bandwidth. Emerging standards such as HSBI aim to push the link rates to 5 Gb/s and higher.

At these data rates, the losses of the channel may be significant. Advanced backplane materials and connectors may be used to mitigate these losses, but system vendors will continue to use traditional FR-4 dielectrics in a significant percentage of high performance backplanes. This choice is driven largely by cost constraints, but reliability and manufacturability issues are also important. Thus,

A. van Roermund et al. (eds.), Analog Circuit Design, 379-396.
© 2003 *Kluwer Academic Publishers. Printed in the Netherlands.*

in addition to driving the channel at high baud rates, the backplane transceivers must also equalize the channel losses. The concentration of transceivers in one system can be on the order of hundreds, due to the frequent need for redundancy and because both ends of the backplane are in the same box. Therefore, the power per Gb/s and area of the transceivers must be kept to an absolute minimum.

2. Backplane Channel

A state-of-the-art high-speed serial backplane may have hundreds of stripline traces with lengths as long as 1 m. Signals are generally routed differentially to minimize crosstalk. In widely-used FR-4 backplanes, the dielectric loss is the dominant term at high frequencies. This phenomenon results in a loss term that scales with frequency and length as:

$$a_d(f) = e^{\alpha_d \ell} = e^{K_d f}, \qquad (1)$$

where ℓ is the length of the line and f is the frequency. The dielectric attenuation constant, α_d, is given by:

$$\alpha_d = \frac{\pi f \sqrt{\varepsilon_r} \tan\delta}{c}, \qquad (2)$$

where c is the speed of light in vacuum, ε_r is the relative permittivity of the dielectric and tan δ is the dielectric loss tangent (tan $\delta \approx 0.018$ in FR-4). The dielectric attenuation constant for FR-4 evaluates to $\alpha_d \approx 3.4$ dB/m at 1 GHz, independent of line geometry or impedance.

Skin effect loss is due to the confinement of the ac current in a conductor to a thin region along its surface. The skin depth in copper is only 2 µm at 1 GHz, compared with a typical trace thickness of 18 µm (0.5 oz), so the ac losses can be considerably larger than the dc resistance of the line would suggest. Due to the high density of connections on a backplane, the line widths are often small (~ 0.2 mm), so that skin effect loss in the conductors is also important. The skin effect loss is a function of the line geometry, and for stripline can be expressed as:

$$a_c(f) = e^{\alpha_c \ell} = e^{K_c \sqrt{f}}, \qquad (3)$$

The skin effect attenuation constant, α_c, is given by [1]:

$$\alpha_c = \frac{2.7\times10^{-3} R_s \varepsilon_r Z_o A}{30\pi(2h-t)}, \quad Z_o < \frac{120}{\sqrt{\varepsilon_r}}, \tag{4}$$

where Z_o is the characteristic impedance of the line, t and h are the thickness and height above the ground plane, respectively, of the line, and A is a geometric factor that includes the effects of line width (w):

$$A = 1 + \frac{2w}{2h-t} + \frac{1}{\pi}\frac{2h+t}{2h-t}\ln\left(\frac{4h-t}{t}\right). \tag{5}$$

The frequency-dependent loss is due to the skin effect resistance, which can be written as:

$$R_s = \frac{1}{\delta\sigma} = \sqrt{\frac{\omega\mu}{2\sigma}}, \tag{6}$$

where ω is the radian frequency and δ is the skin depth in a conductor having conductivity σ and permeability μ.

A 200 μm wide copper stripline with 50 Ω impedance would have a skin effect attenuation constant of $\alpha_c \approx 2.9$ dB/m at 1 GHz. The skin effect loss dominates

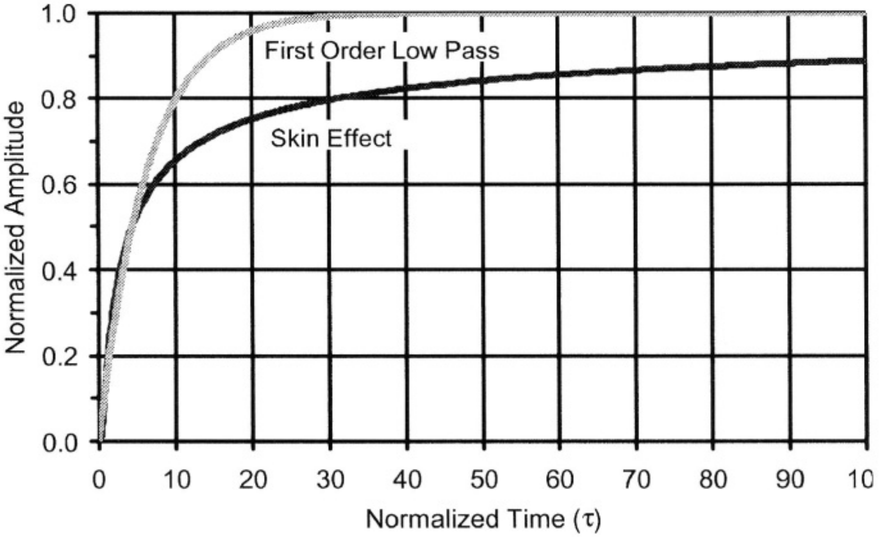


Fig. 1. Step responses of skin effect channel and first-order low pass channel (scaled to equal delay to 50%).

at low frequencies, since it increases as the square root of frequency, while the dielectric loss increases linearly with frequency. The crossover is in the low GHz range for practical backplane line aspect ratios.

An example of the step response through a channel with predominantly skin effect loss is shown in Figure 1 [2]. The skin effect loss gives a long transient tail, which can cause significant inter-symbol interference for uncoded bit sequences with long run lengths and bit periods comparable to the channel rise time.

The connectors that mate the line cards to the backplane contribute to the channel loss by causing impedance mismatches. The connectors are also a primary source of crosstalk. Fortunately, low-loss materials and higher-fidelity connectors are available to improve the channel loss and crosstalk. Table 1 summarizes some of the important properties of a range of a few backplane dielectric materials.

Figure 2 compares the frequency response of a legacy FR-4 backplane with conventional VHDM-HSD connectors to that of a backplane implemented in Rogers 4350, with more advanced GbX connectors. Both backplane traces are 1 meter long. Note that the low-loss backplane has higher ripple since the connector reflections are attenuated less.

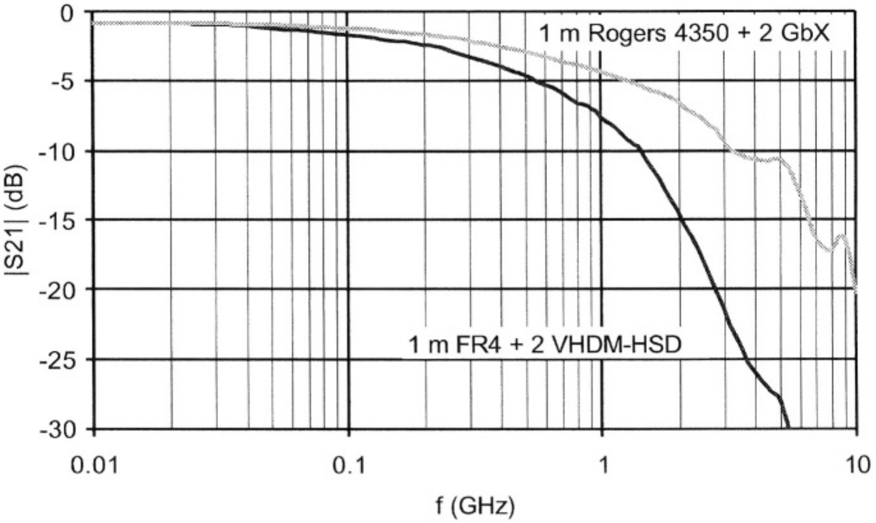


Fig. 2. Simulated frequency responses of 1 meter FR-4 and Rogers backplanes with two connectors.

Material	ε_r	tan δ	Relative Cost
FR-4	4.4	0.018	1.0
Getek	3.9	0.012	1.1
Rogers 4350	3.5	0.004	2.1
ARLON CLTE	2.9	0.0025	6.8

Table 1. Backplane dielectric material properties.

3. Binary Transceivers

Serial backplane transceivers predominantly use binary (two-level PAM) signaling because its relatively low implementation complexity allows for very high speed of operation, commensurate with the multi-Gb/s data rates needed. Clocks are generally embedded with data to minimize the impact of path skew. For flexibility and expandability, the data processing in a generic network system is generally partitioned into N line cards connected through a backplane. These line cards may be interconnected with dedicated links, but this requires $O(N^2)$ links, some of which may traverse the whole backplane. The link count can be reduced to $O(N)$ if all links are routed through a central switch card; if the switch card can be placed at the center of the chassis, the maximum link length is approximately halved. However, this switched backplane approach exacerbates the power and congestion problems on the switch card and the ICs that implement the switching function. Thus, the backplane transceivers integrated in the switch chips often have the most severe power and integration requirements.

The simplest and lowest power means of implementing the switch function is with an asynchronous switch. An asynchronous switch does not enforce any data rate or timing relationships between the I/Os, which allows different data rates to be passed on different I/Os. Since there is generally dead time when the switch reconfigures, this type of switch is better suited for circuit switching rather than packet switching applications.

Figure 3 shows a representative 34 input × 34 output asynchronous crosspoint switch. It uses pre-amplifiers and unclocked comparators for high input sensitivity. The signal path is fully differential and based on current mode logic (CML). The output drivers are differential open collector current switches. The output level is digitally programmable on a per-output basis to minimize power and crosstalk. Trimmed on-chip terminations minimize reflections at package

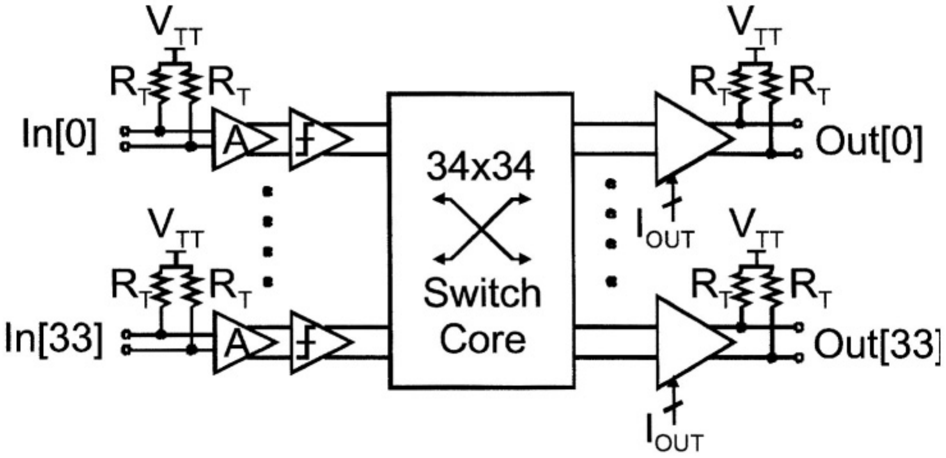


Fig. 3. Block diagram of a 34x × 34 asynchronous crosspoint switch with on-chp terminations and programmable outplut level.

interfaces. This circuit is implemented in 0.35 μm BiCMOS and can achieve 40 ps p-p jitter at 3.2 Gb/s with less than 60 mW/channel [3]. Other asynchronous switches have reached port densities of 160 × 160 and I/O and data rates up to 12.5 Gb/s, albeit at much higher power [4].

The major drawback of a purely asynchronous transceiver is the accumulation of timing errors, such as jitter and duty cycle distortion. These can be controlled in the IC itself, but the low pass backplane channel introduces significant pattern-dependent jitter. These effects are illustrated in Figure 4, which shows the measured eye diagrams at 3.2 Gb/s on several points of a 1 m long FR-4 backplane with a total of four high-density connectors. The differential transmitter embedded in an asynchronous crosspoint switch drives the input to the left half of the backplane. The received eye at the midpoint of the first backplane already shows reduced eye height, increased jitter and degraded rise time. An asynchronous transceiver at the second switch at the midpoint of the backplane can restore the vertical eye opening, but adds to the overall jitter. Thejitter and eye height of the signal at the end of the right half of the backplane are marginal. These measurements show that a purely asynchronous transceiver without equalization can deliver 3.2 Gb/s data over more than 50 cm of conventional backplane. Transceivers designed for higher data rates and trace lengths where losses are higher may incorporate equalization and clock and data recovery circuits to compensate for channel loss and its associated pattern-dependent jitter.

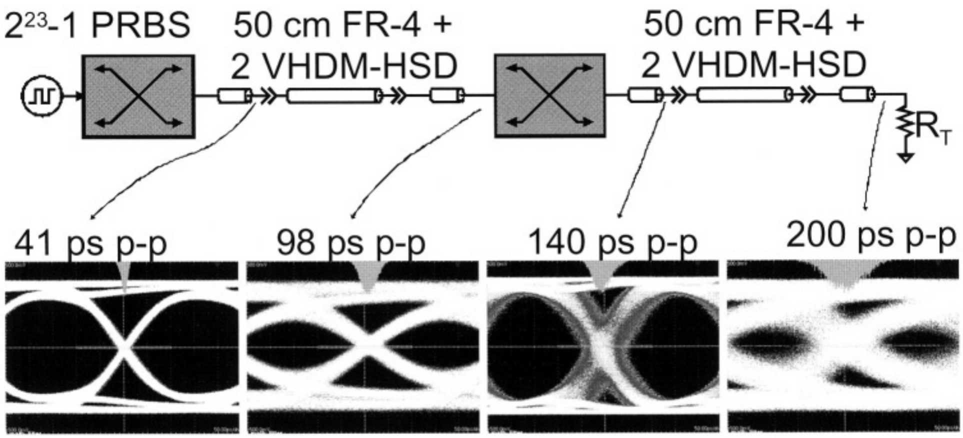


Fig. 4. Measured 3.2 Gb/s at several points along an FR-4 backplane with asynchronous transceivers.

4. Equalization

In principle, the backplane loss can be compensated with a filter having the inverse of its frequency response. This filter can be implemented in the transmitter or receiver and may be either discrete time or continuous time. A continuous time receive equalizer can be relatively easy to implement since it does not require a recovered clock. Since the skin effect and dielectric losses are proportional to line length, a variable filter equalization filter should have a gain in dB of the form:

$$H_{dB}(\omega) = 20\log_{10}\left|H(\omega)\right| = \beta(\ell)G_{max}(\omega), \quad 0 < \beta(\ell) \le 1, \qquad (7)$$

where $G_{max}(\omega)$ is the reciprocal of the maximum channel loss (vs. frequency) and $\beta(\ell)$ is a scaling parameter proportional to trace length. This parameter may be programmed through an external digital interface or set adaptively via a circuit that minimizes some receiver error metric (e.g. deviation in received rms power from a target level).

A Bode equalizer [5] approximates the desired behavior of Eqn. 7 as a rational transfer function with variable coefficients. It can be expressed as:

$$H_{Bode}(\omega) = \frac{(1-b) + bG_{max}(\omega)}{b + (1-b)G_{max}(\omega)}, \quad 0 < b \le 1, \qquad (8)$$

where b is another length scaling parameter. This form gives a flat frequency response for $b = 0.5$, gain for larger values of b, and loss for smaller values of b. A more readily realizable form that gives similar results for the high gain ($b \sim 1$) case is:

$$H_f(\omega) = (1-a) + aG_{max}(\omega), \quad 0 < a \leq 1, \tag{9}$$

where a is yet another scaling parameter.

Figure 5 illustrates how an equalizer of the form of Eqn. 9 can be implemented. In the block diagram shown in Fig. 5a, the output is generated by summing a unity-gain path with a high pass path with complementary weights, ($1 - a$) and a, respectively. Figure 5b is a simplified differential circuit implementation. Transistors $\mathbf{M}_{11}$ and $\mathbf{M}_{12}$ and the 2R degeneration resistor form a linear

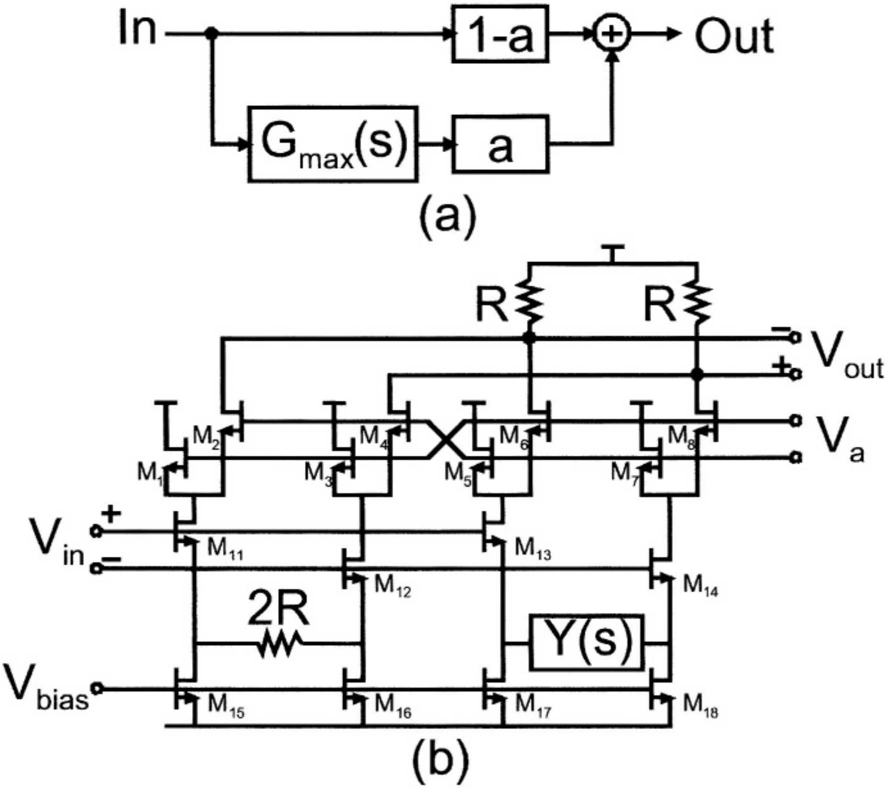

In
1-a
+
Out
$G_{max}(s)$
a
(a)
R
R
V_{out}
−
+
V_a
M_1
M_2
M_3
M_4
M_5
M_6
M_7
M_8
V_{in}
+
−
M_{11}
M_{12}
M_{13}
M_{14}
2R
Y(s)
V_{bias}
M_{15}
M_{16}
M_{17}
M_{18}
(b)

Fig. 5. Continuous time receive equalizer. a) Block diagram. b) Simplified schematic.

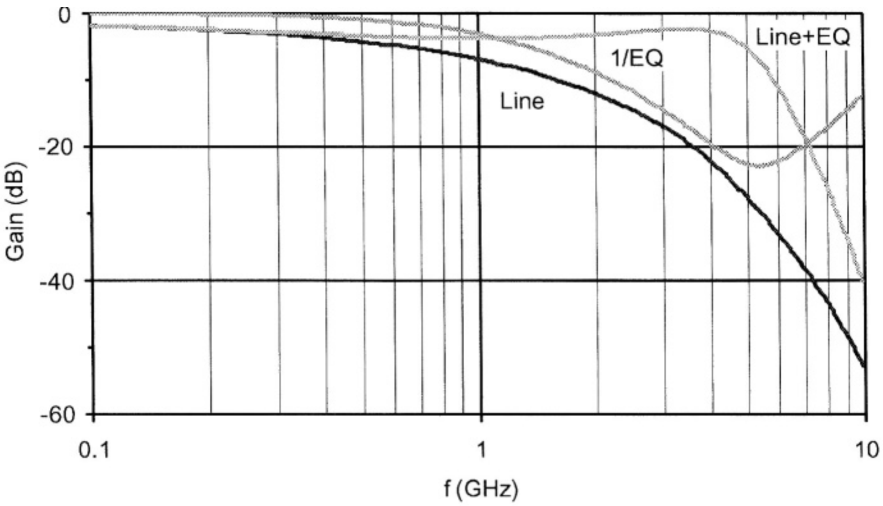


Fig 6. Frequency response of a 1 m FR-4 backplane with and without a continuous time receive equalizer.

transconductor. Transistors M_{13} and M_{14} and the Y(s) admittance network form a transadmittance stage. Four differential pairs (M_1/M_2, M_3/M_4, M_5/M_6 and M_7/M_8) driven by control voltage V_a weight the output currents from the transconductance and transadmittance stages. The admittance network is related to the desired maximum equalization gain curve by the following expression:

$$Y(s) = \frac{G_{max}(s)}{2R}. \tag{10}$$

It can be implemented as a simple R-C network for gain slopes vs. frequency of up to 20 dB/decade. Multiple equalizer stages of this form may be cascaded to achieve higher gain slopes.

Figure 6 shows the effectiveness of this type of equalizer in the frequency domain. The frequency response of a 1 m FR-4 backplane with benign connectors follows closely the inverse of the frequency response of a programmable continuous-time receive equalizer for frequencies up to about 5 GHz. Thus, the frequency response of the cascade of the channel and the equalizer has a relatively flat gain out to this frequency. The equalizer in this example is implemented in a 30 GHz f_T silicon BiCMOS process and is programmed to minimize the jitter at 6.25 Gb/s.

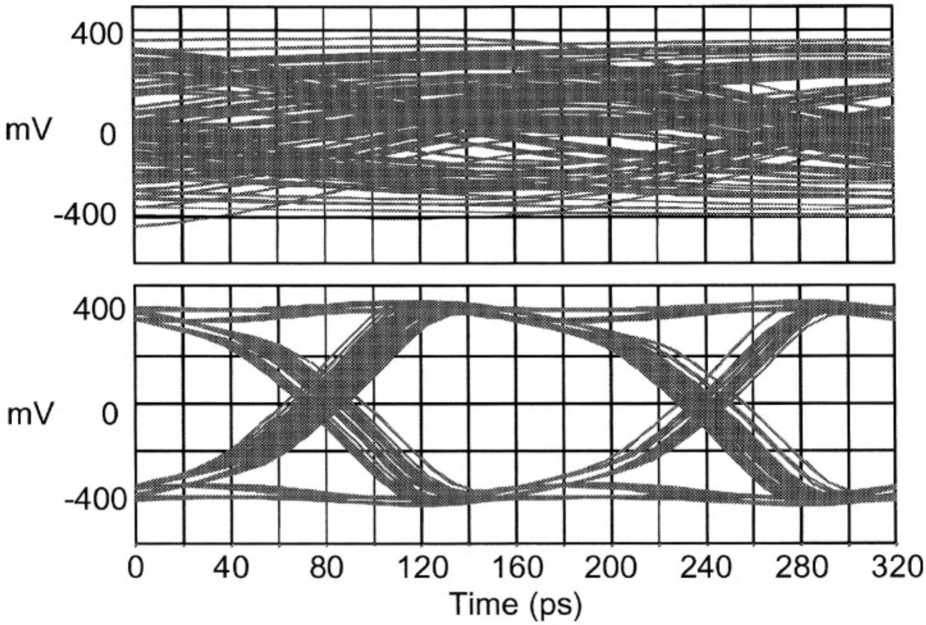


Fig. 7. Equalization of a 1 m FR-4 backplane with a continuous time receive equalizer. Top: received eye diagram at 6.25 Gb/s NRZ; bottom: output of the equalizer.

Figure 7 illustrates the performance of this continuous time receive equalizer in the time domain. The upper eye pattern is the received eye before equalization and the lower eye is the output of the equalizer before slicing. The equalized jitter is approximately 0.2 unit interval (bit period).

Equalization can also be done in the discrete time domain using finite impulse response filters. The approach makes sense in a synchronous transmitter where the input data is already sampled and quantized. This technique is often referred to as transmit pre-emphasis. Figure 8 shows a three-tap FIR transmit filter that realizes the cascade of two real zeros as well as the response of this filter to a bit sequence. Note that the amplitude in the vicinity of data transitions is accentuated, which pre-compensates for the low pass response of the backplane.

Figure 9 shows this type of 3-tap FIR transmit filter merged with a differential binary transmitter. The delays are implemented with two flip flops. The raw and delayed input bit streams are presented to the inputs of a bank of one-bit, current output DACs with variable output currents. The output currents are set proportional to the desired tap weights. The summation is done in current mode at the common outputs of the drivers. The driver segments can be implemented

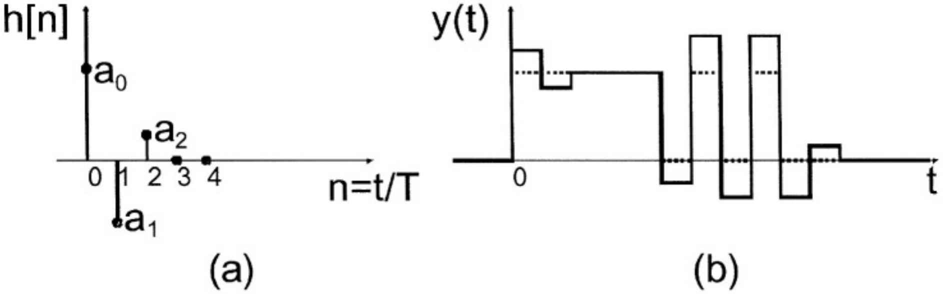


Fig. 8. a) Impulse response of 3-tap transmit FIR filter. b) Time domain response to arbitrary bit sequence.

as differential current switches with quasi-DC programmable tail current sources. The pre-driver is implemented as a low-swing differential circuit for low noise generation and high supply noise immunity.

The behavior of the FIR transmit filter can be evaluated in the frequency domain using Z-transform methods. We will consider the simplest case of a two-tap FIR filter (one tap of pre-emphasis); this analysis can be readily extended to longer filter impulse responses. The Z domain transfer function of a two-tap transmit filter may be written as:

$$H(z) = a_0 + a_1 z^{-1} = a_0(1 - rz^{-1}), \qquad (11)$$

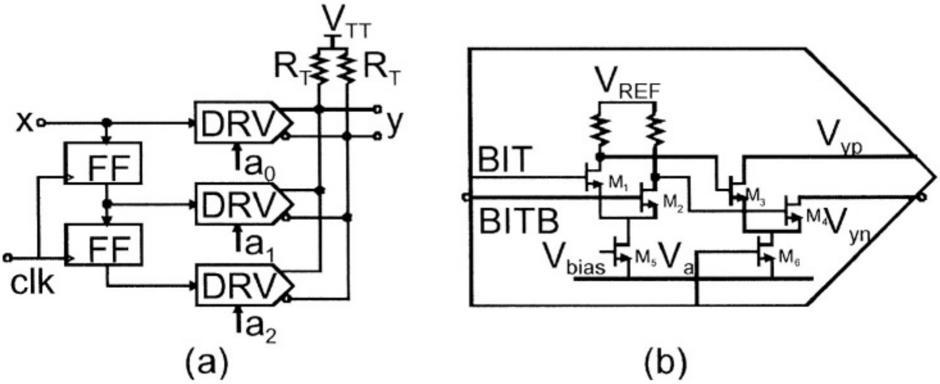


Fig. 9. a) A differential backplane transmitter with binary signaling and an integrated 3-tap transmit filter. b) Detail of a driver segment.

where a_0 and a_1 are the tap weights, z^{-1} is the unit sample delay (equal to the bit period in this case), and $r = -a_1/a_2$ is relative pre-emphasis weight. To implement a high-pass function $a_1 > 0$ and $a_1 < 0$, which implies $r > 0$. The frequency response of this transfer function is:

$$H(\omega) = a_0(1 - r \cdot e^{j\omega T}), \tag{12}$$

where T is the bit period and j is $\sqrt{-1}$. The magnitude of the frequency response is given by:

$$|H(\omega)| = a_0\sqrt{1 + r^2 - 2r\cos\omega T}, \tag{13}$$

which has a minimum at dc:

$$|H(\omega = 0)| = a_0(1 - r) = a_0 + a_1, \tag{14}$$

and a maximum at integer multiples of half the bit rate:

$$|H(\omega T = n\pi)| = a_0(1 + r) = a_0 - a_1. \tag{15}$$

where a_0 and a_1 are the tap weights, z^{-1} is the unit sample delay (equal to the bit period in this case), and $r = -a_1/a_2$ is relative pre-emphasis weight. To implement a high-pass function $a_1 > 0$ and $a_1 < 0$, which implies $r > 0$. The frequency response of this transfer function is:

$$H(\omega) = a_0(1 - r \cdot e^{j\omega T}), \tag{12}$$

where T is the bit period and j is $\sqrt{-1}$. The magnitude of the frequency response is given by:

$$|H(\omega)| = a_0\sqrt{1 + r^2 - 2r\cos\omega T}, \tag{13}$$

which has a minimum at dc:

$$|H(\omega = 0)| = a_0(1 - r) = a_0 + a_1, \tag{14}$$

and a maximum at integer multiples of half the bit rate:

$$|H(\omega T = n\pi)| = a_0(1 + r) = a_0 - a_1. \tag{15}$$

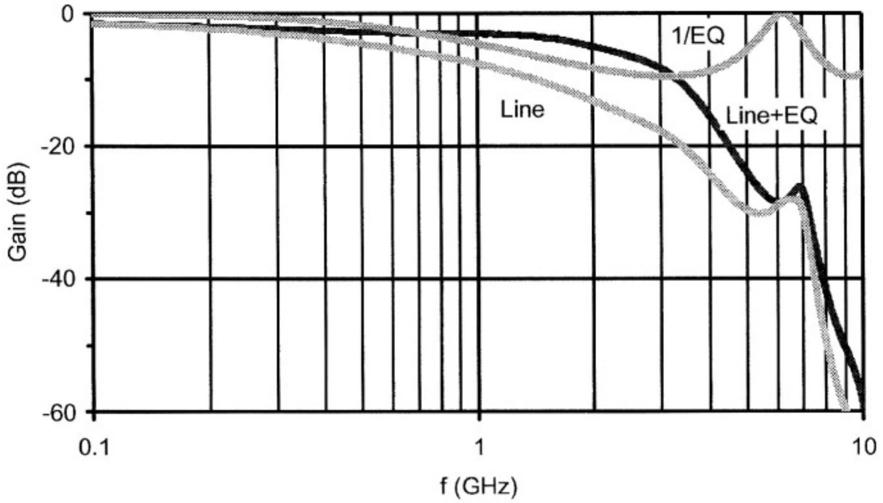


Fig 10. Frequency response of a 1 m FR-4 backplane with and without a two-tap FIR transmit equalizer.

The maximum boost is the ratio of these quantities:

$$Boost = \frac{|H(\omega T = n\pi)|}{|H(\omega = 0)|} = \frac{1+r}{1-r} = \frac{a_0 - a_1}{a_0 + a_1}. \tag{16}$$

If the transmitter is power-constrained or swing-limited, the sum $|a_0| + |a_1|$ is also bounded. Since $a_1 < 0$ for the high-pass case, this constraint implies that to achieve boost, the low frequency components are effectively attenuated (de-emphasized) with respect to the high frequency components. For a receiver with finite sensitivity, this effect limits the maximum pre-emphasis boost. Mismatches in the transmitter can also place limitations on the maximum usable boost. Fortunately, modest boost levels on the order of 10 dB ($r \approx 0.5$) can have a dramatic impact on eye opening.

Figure 10 shows the frequency response of 1 meter of FR4 with two GbX connectors, the inverse of the 2-tap FIR transmit equalizer frequency response (neglecting the effects of the implicit zero-order hold in the transmitter), and the composite frequency response of the channel and transmit equalizer. The bit period is $T = 160$ ps and the tap weights are set for 9.5 dB of peak boost ($r = 0.5$) at the Nyquist rate of 3.1 GHz. The unequalized backplane half-power bandwidth of 500 MHz is increased to 1.9 GHz after equalization.

Figure 11 shows the simulated time domain performance of this 2-tap FIR transmit equalizer over the same 1 m FR-4 channel as Figure 10. The 50 Ω backplane traces are doubly terminated. The upper eye diagram shows the far end of the line with no equalization; the eye is barely distinguishable. The lower eye diagram shows the received signal with the transmit equalizer enabled (a_0 = 16 mA, a_1 = −8 mA, r = 0.5). The equalized jitter is approximately 0.2 unit interval (or bit period). The vertical eye height is improved to 190 mV and the peak-to-peak jitter is ~ 0.19 UI. Note that the low-frequency amplitude is lower than in the unequalized case.

5. Multi-Level Transceivers

When the loss slope of the channel becomes excessive, it may become advantageous to encode multiple bits per symbol to reduce the baud rate [6]. The loss and crosstalk at the reduced Nyquist frequency would have to be

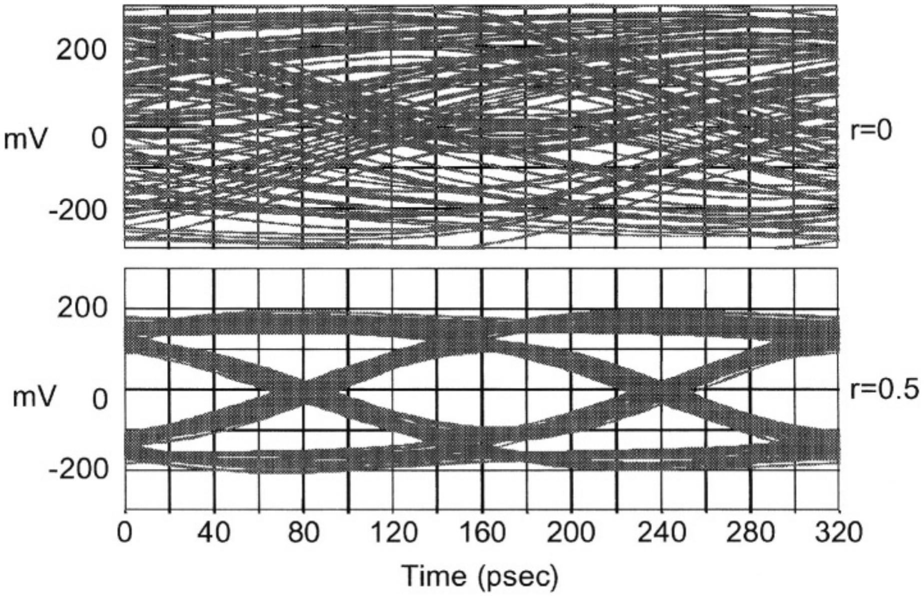


Fig. 11. Equalization of a 1 m FR-4 backplane with a continuous time 2-tap FIR transmit equalizer. Top: received eye diagram at 6.25 Gb/s NRZ with equalizer disabled; bottom: received eye diagram with equalizer enabled.

sufficiently lower to compensate for the reduced symbol spacing resulting from the higher number of levels needed to encode multiple bits per symbol. For example, by moving from binary or two level pulse amplitude modulations (PAM 2) to four level pulse amplitude modulation (PAM 4), the baud rate could be halved, but the symbol spacing would be reduced by a factor of three if the transmit power is limited.

Figure 12 illustrates this tradeoff for a specific channel frequency response corresponding to a 1 m FR-4 backplane with low loss connectors. A simple metric to compare alternative coding schemes is to evaluate the ratio of the signal amplitude to the sum of the impairments at the receive end, The upper two curves in Figure 12 (labeled N=2 and N=4) represent the effects of channel loss on the received signal. The binary signal swing is assumed to be 1 V p-p. The multi-level curve (N=4) is shifted down from the upper curve by 9.5 dB to account for the 3x reduction in transmit symbol spacing due to the higher number of quantization levels. The impairments are split into two components, indicated as near end crosstalk (NEXT) and receiver (RX) sensitivity in the graph. The NEXT curve follows the envelope of the measured all-hostile crosstalk for a VHDM-HSD connector with the two worst aggressors removed (assuming some optimization is done in the connector pin assignments to avoid this worst case). The receiver sensitivity aggregates the effects of noise, offset, on-chip interference & finite gain in the receiver. In this example, it is taken to be about 30 mV p-p and dominates the NEXT.

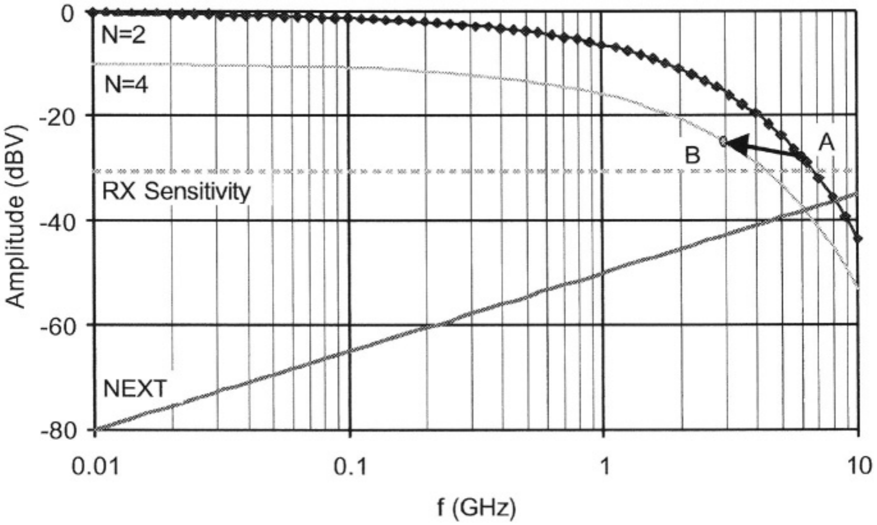


Fig. 12. Signal to (noise + interference) vs. frequency of binary (PAM-2) and multi-level (PAM-4) signaling at equal bit rates.

Point A in Figure 12 corresponds to a 12 Gsymbol/sec baud rate using binary signaling, giving 12 Gb/s. Point B represents 6 Gsymbol/sec baud rate using PAM-4 multi-level signaling and also yielding 12 Gb/s. In this case, moving to the more spectrally efficient (N=4) multi-level scheme results in a modest signal to impairment advantage. Using this metric, it can be shown analytically that for a channel whose loss is dominated by skin effect loss (i.e. proportional to the square root of frequency), PAM-4 becomes favorable over binary where the loss at the Nyquist rate is greater than 33 dB using binary signaling. If the channel loss is dominated by dielectric loss, the break-even point is 19 dB of loss at the Nyquist rate.

When comparing binary and multi-level transceivers, the cost of implementation must also be considered. Figure 13 compares block diagrams of binary receivers and transmitters to their multi-level (W-bit per symbol) counterparts. The power and area of the multi-bit receiver are similar to those of a low-resolution A/D converter and can be expected to grow between $O(N)$ for a fully pipelined architecture and order $O(2^N)$ for a flash implementation. The multi-level transceiver also incurs the cost of a reference generator. The multi-level transmitter complexity should scale in a manner similar to that of its receiver depending on the level of segmentation chosen in the transmit DAC. However,

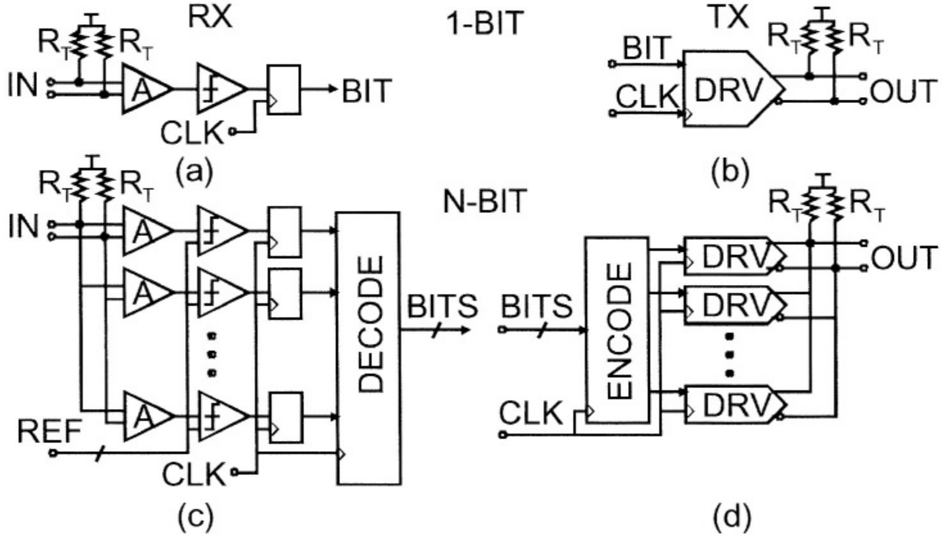


Fig. 13. Binary vs. multi-level circuit complexity. a) Binary receiver. b) Binary transmitter. c) Multi-level receiver. d) Multi-level transmitter.

the power in an efficient transmitter can be dominated by the output current needed to swing the load, so the total power multi-level transmitter should grow more slowly than the complexity. Multi-level transceivers reported to date have been much higher power than their binary counterparts, which makes them less suitable for high-density applications.

6. Conclusion

Backplanes pose significant problems to the transmission of high bit-rate serial digital signals. Equalization in either the transmitter or receiver can compensate for the channel loss while maintaining compatibility with conventional binary transceivers. Multi-level transceivers can reduce the symbol rate to frequencies where the loss is lower. The overall power, area and complexity of the transceivers are important since many applications require the integration of many channels per chip.

References

[1] C. Nguyen, *Analysis Methods for RF, Microwave, and Millimeter-Wave Planar Transmission Line Structures,* John Wiley, New York.

[2] R.B. Adler, L.J. Chu, *Electromagnetic Energy Transmission and Radiation,* John Wiley, New York.

[3] AD8152 datasheet, Analog Devices.

[4] H. Veenstra et al., "A 20-Input 20-Output 12.5 Gb/s Cross-Point Switch with Less Than 2ps RMS Jitter," ISSCC Digest of Technical Papers 46 (2003), pp. 174-175.

[5] H. Bode, "Variable Equalizers," *Bell Syst. Tech J.,* Apr. 1938.

[6] W. Ellersick, *et. al.,* "A Serial Link Transceiver Based on 8 GSa/s A/D and D/A Converters in 0.25 μm CMOS," *ISSCC* 2001.